Fast Relevant Simulation in Finance

by

Kazuhiro Iwasawa

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics

New York University

Approved: ______________________________

Jonathan Goodman

Advisor

Acknowledgement

I would like to thank Professor Jonathan Goodman for his wonderful advice on both mathematics and life in general. I learned so much from him. I appreciate all the energy, time, and insights he gave to me about scientific computing. I am grateful to Professor Varadhan for his excellent advice on the Large Deviation Theory and the geometric ergodicity in general. I would like thank Professors Marco Avellaneda, David Cai, and Robert Kohn for their excellent advices on the topic. I am also grateful to Professor Bernard Delyon (Universite de Rennes) for his excellent lecture notes and comments via email communications. I would like to thank RBS Greenwich Capital, Inc. for accommodating me through the period I am writing my dissertation. I would also like to thank my wife and my parents for supporting me throughout my studies. Finally, I thank all the Courant staff for supporting me for my studies at Courant.

Preface

This paper introduces new Monte Carlo methods for computing Value-at-Risk(VAR) in finance. Monte Carlo methods will be classified into two major categories - i.i.d. Monte Carlo and Markov Chain Monte Carlo. This paper proposes new techniques for each category. They are motivated by the Large Deviation Theory concerning rare event simulation.

Chapter 1 is an introduction to the concept of VAR, and simulation methods in general. A brief introduction to the Markov Chain Monte Carlo will also be given. Chapter 2 introduces importance sampling, and the control variate method. The former method can be generalized to Markov Chain Monte Carlo case as well. Then their typical use for VAR as well as pros and cons will be analyzed. Chapter 3 introduces a new technique for an i.i.d. Monte Carlo method. Most Monte Carlo simulations fall into this category. It is based on i.i.d. sampled random numbers. The new technique is based on the combination of importance sampling, non-linear optimization, and newly proposed exponential twisting density. The theoretical justification will be given based on the Large Deviation Theory and the Laplace method. The numerical results thereof will be given in chapter 4.

Chapter 5 introduces Markov Chain Monte Carlo, and new technique based on the Metropolis $M(RT^2)$-within-Gibbs algorithm combined with the Robbins-Monro algorithm from stochastic approximation theory. Chapter 6 contains series of theoretical analysis and proofs. Chapter 7 defines various path-dependent options, and proposes the use of MCMC for computing VAR. The numerical results will be given in chapter 8. The conclusion of this paper is given in chapter 9. This paper contains a series of appendices at the end followed by bibliography.

Abstract

Fast Relevant Simulation in Finance

Author: Kazuhiro Iwasawa

Thesis Advisor: Jonathan Goodman

This paper proposes effective simulation methods for sampling rate events in finance. In particular, the Monte Carlo method will be improved algorithmically to compute Value-at-Risk (VAR) for various portfolios containing non-linear derivatives such as options (including path-dependent options). We propose various variance reduction techniques motivated by the Large Deviation Theory for rare event simulation. The most common variance reduction technique for this is called importance sampling.

At first, the independent sampling (i.i.d.) case will be analyzed. In this paper, non-linear optimization will be used to explicitly compute an importance sampling density to estimate VAR for portfolios containing European options. Then a more general approach will be used to perform complex simulation using Markov Chain Monte Carlo ('MCMC' for short). We will introduce new technique based on the Metropolis-Hastings $M(RT)^2$ algorithm. This technique will effectively sample from an importance twisting density which is being optimized at the same time using the Robbins-Monro algorithm from stochastic approximation theory. It turns out that this method can be used to compute VAR containing path-dependent options as well as high dimensional VAR very effectively.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

1.1 Value-at-Risk (VAR)

Value-at-Risk (VAR) has become a major concept in the analysis of market risk in finance ([21],[60]). Typically, the value of a portfolio in the future is a random variable. Let us define the future portfolio value as V, and the initial portfolio value today as v_o. Then for a given confidence interval p, VAR is defined as a real number [1] $v_p - v_o$ satisfying:

$$P(V \leq v_p) = p \tag{1.1}$$

It is basically the maximum possible loss a firm (or desk, etc) will lose in a given time horizon with certain confidence p. The popularity of VAR lies in the simplicity of expressing firm's risk exposure in one number. Typically the 1-day 95 % VAR ($p = 0.05$) and the 10-days 99 % VAR ($p = 0.01$) are used in the banking industry. Recently BIS (Bank for International Settlements) have made regulatory requirement to use the latter for capital allocation purposes.

Computing VAR involves estimating the very small probability p for the tail event of the distribution of V in the loss region. There are various methods for computing VAR ([21]) ranging from analytic methods to simulation methods. Almost all analytic methods ([60], [8]) lose accuracy in the tail (or 'extreme quantile', $p < 0.05$) or if the portfolio function V is distributed in highly non-gaussian way. The alternative is to use simulation methods for the extreme quantile cases using Monte Carlo.

Conventional Monte Carlo loses accuracy for a small p. We seek Monte Carlo methods that give better accuracy for small p. This is related to the general area of Large Deviation Theory and rare event simulation. Rare event simulations are often improved by using importance sampling (defined in the next chapter). There are several importance sampling methods using analytic formulae as guidance ([36], [38], [39], [40], [42]). Though they work for some

[1] $V - v_o$ is the P&L (profit or loss from today to sometime in the future)

extreme quantile cases, they are known to fail if the portfolio V is distributed in highly non-gaussian way, or there is more than one way to reach the VAR region (more than one local minimum).

We investigate importance sampling techniques with less dependency on analytic methods. Still motivated by the Large Deviation Theory, we will propose new alternative Monte Carlo methods for computing VAR. These methods can even be applied to compute VAR for portfolio containing path-dependent options. We shall give brief descriptions of our methods.

1.2 I.I.D. Monte Carlo and Markov Chain Monte Carlo Strategies

We examine two rare event importance sampling strategies. The first method uses a simple reweighting intended to produce sampling near the most likely scenario in the VAR region. This strategy is appropriate for portfolios that are not path dependent, that is, for $V = V(x)$ with $x \in R^d$ with a simple explicit density $f(x)$. Portfolios containing underliers and European style options or American options that do not get exercised in the VAR period have this property. This method uses non-linear optimization to identity $\beta \in B_v$ such that $f(\beta) = \max_{x \in B_v} f(x)$ where $B_v \equiv \{x : V(x) \leq v\}$. β is called the 'local minimum', and B_v is called the 'VAR region'. We then propose a new exponential twisted density along a direction perpendicular to the VAR surface $V(x) = v$ near the local minimum. This density is quite simple and can be sampled very easily. For some portfolios, there can be many such β. We call these the 'multiple local minima'. For these cases, the twisted density will be a convex combination of the above twist for each β. The details of this technique are explained in chapter 3.

The second method is more complicated and is intended for portfolios containing path-dependent options or other situations where explicit probability maximization is infeasible. Here we propose to sample from the twisted density:

$$\pi_\theta(x) = f(x) \cdot e^{-\theta \cdot max(V(x), v)} \tag{1.2}$$

where f is the original correlated normal density and θ is chosen adaptively to produce samples near the VAR boundary $V = v$. In this case, we are unable to produce independent sampling of the density π_θ. Therefore, we use Markov Chain Monte Carlo. We implement multidimensional Metropolis Hasting $M(RT)^2$ algorithm to sample from π_θ, and use the Robbins-Monro algorithm from stochastic approximation theory to update θ. The characteristic of this method is that this formula (1.2) peaks at almost all local minima, thus more sampling are likely in the VAR region. The details of this method will be shown in the chapter 5. We will show that this algorithm works from theoretical view in chapter 6.

In the next two sections, the basic Monte Carlo Methods for computing VAR, and basic mathematical finance terminologies and notations will be explained. Those who are familiar with the general practice of Monte Carlo VAR calculation, and mathematical finance in general can skip these section. In chapter 2, we describe general variance reduction techniques - importance sampling, and control variates. We have included the latter because we used this technique for relatively non-extreme quantile VAR calculation for testing purposes. For those who are familiar with general importance sampling techniques can skip the entire chapter 2 as well.

1.3 Basic Monte Carlo Method for VAR

In general, the usual task is to compute v for given p. Typical Monte Carlo requires generating [1] N instances of random numbers [2] V, and sort these values of V in ascending order. The $N \cdot p$-th value is the estimate for v_p. The error analysis for this problem can be computed by using order statistics.

A simpler but also challenging problem is to estimate the distribution function $F(v) = P(V \leq v)$ for v in the tail so that F is small. This also involves rare event simulation. Recovering v from F is straightforward [3]. Therefore we shall focus on estimating $F(v) = p$ for a given v.

Let us assume that X has a probability density $f \equiv f(x), x \in R^d$. Let us further define the set $B_v = \{x : V \equiv V(x) \leq v\}$. Then p can be expressed as:

$$p = P(V \leq v) = \int_{R^d} I_{B_v}(x) f(x)\, dx \tag{1.3}$$

where I is a characteristic function. In Vanilla Monte Carlo simulation, its estimator is given by:

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} I_{B_v}(X_i) \tag{1.4}$$

If the sampling is done i.i.d. (independently identically distributed;), then the law of large numbers guarantees $\hat{p} \to p$ and the Central Limit Theorem (CLT) guarantees that its error goes down by $\frac{\sigma}{\sqrt{N}}$ where σ is a standard deviation given by: $\sigma = \sqrt{p(1-p)}$. Typically the statistical error is given by $\varepsilon = \frac{2\sigma}{\sqrt{N}}$ which is an error bound with 95 % confidence. The relative error is given by $\frac{\varepsilon}{p}$. For rare event simulation, p is small. Therefore, the relative error is approximated as $\frac{\varepsilon}{p} = \frac{2\sqrt{p(1-p)}}{\sqrt{N}p} \approx \frac{2}{\sqrt{Np}}$. This means that as p becomes smaller, the relative error increases. So, we have to increase N to keep the relative error

[1] Typically $N = 10{,}000$ to a million.
[2] Section 1.4 for the details of how to generate V.
[3] Section 2.1 for the details

small when p is very small. This could be computationally very expensive for many portfolios. Variance reduction techniques result in small variance. General variance reduction techniques will be explained in chapter 2.

1.4 Basic Mathematical Finance Terminologies and Notations

In the previous sections, a portfolio value V was defined as a random number in R^1. However; in general, V is a function of several, often correlated underlying risk factors $S = (S_1, S_2, ..., S_d)$. Therefore, a portfolio value of a firm (or, desk, trader, etc) V can be re-written $V \equiv V(S)$ where $S(t) = (S_1, S_2, ..., S_d)$ is a vector of d dimensional underlying risk factors. A portfolio $V(S)$ can consist of equities, foreign exchange products, fixed income securities such as government bonds/corporate bonds/swaps/eurodollar future contracts, commodities, or options on these products. A vector of underlying risk factors S can be equity prices, foreign exchange rates, interest rates, etc. Let us assume that the underlying risk factors S_i are d dimensional correlated lognormal random processes (Some underliers are modeled to follow normal, or other processes. We do not consider these cases here since the results of this paper can be easily generalized to these cases) where each underlying risk factor S_i is driven by the following SDE (stochastic differential equation) :

$$\frac{dS_i}{S_i} = \mu_i dt + \sigma_i dW_t^i, i = 1, 2, ...d \tag{1.5}$$

where μ_i is a drift, σ_i is a volatility, and W_t^i is a Weiner process for each underlying S_i. The above SDE admits following solution ([48]):

$$S_i = S_i(0) e^{\sigma_i W_t^i + (\mu_i - \frac{1}{2}\sigma_i^2)t}, i = 1, 2, ...d \tag{1.6}$$

This formula also holds also if the process W_t^i are correlated. In that case, we define Σ_X to be the correlation matrix, which is also the covariance matrix for W_t^i with $t = 1$.

In general, one generates random numbers V using Monte Carlo as follows. At first, perform the Cholesky decomposition of Σ_X as $\Sigma_X = C_X^T \cdot C_X$ once. Next, generate d independent standard normal random numbers $\hat{X}$. Then compute correlated d normal random numbers X by the vector multiplication as $X = C_X^T \cdot \hat{X}$. $W = X\sqrt{t}$ are plugged into the formula (1.6) . Finally use these S to evaluate $V(S)$. Therefore, if N iterations of Monte Carlo simulations are requited, just repeat the above steps N times (except the Cholesky decomposition of Σ_X). Since $V(S)$ is also a function of x, and our techniques are applied to Gaussian random numbers x, we shall refer V as $V(x)$ throughout this paper.

Chapter 2

General Variance Reduction Techniques for Monte Carlo

2.1 Importance Sampling Method for VAR calculation

Importance sampling is a variance reduction technique which focuses on finding right density for the integrand so that more samples are done near the region of interest. The trick comes from multiplying the integrand with an appropriate function called the importance function (or sometimes called the likelihood ratio) so that the integration with the new density (called the twisted density) is preserved mathematically. Let us describe this technique in more detail.

Our objective is compute VAR probability $p = \int_{R^d} I_{B_v}(x)f(x)dx$ where $B_v = \{x : V(x) \leq v, x \in R^d\}$, I is a characteristic function, and f is a probability density. The importance sampling technique calls for introducing a new twisted density π where f is absolutely continuous with respect to π so that the probability can be expressed as:

$$\begin{aligned} p = P(V \leq v) &= \int_{R^d} I_{B_v}(x)f(x)\,dx \\ &= \int_{R^d} I_{B_v}(x)\frac{f(x)}{\pi(x)}\pi(x)\,dx \end{aligned} \tag{2.1}$$

The term $\frac{f(x)}{\pi(x)}$ is called the likelihood ratio. This process is also called the change of measure. The estimator for p under π is given as:

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} I_{B_v}(X_i)\frac{f(X_i)}{\pi(X_i)} \tag{2.2}$$

where N is the number of iterations. On the other hand, when π is known only up to a constant (mostly the case in MCMC), then π is expressed as $\pi = C \cdot \bar{\pi}$

where C is an unknown constant. Since $\int_{R^d} \frac{f(x)}{C\cdot\bar{\pi}(x)}\pi(x)\,dx = 1$, the estimator for $\hat{p}$ can be given as:

$$\hat{p} \quad = \quad \frac{\frac{1}{N}\sum_{i=1}^{N} I_{B_v}(X_i)\frac{f(X_i)}{\pi(X_i)}}{\frac{1}{N}\sum_{i=1}^{N}\frac{f(X_i)}{\pi(X_i)}} \tag{2.3}$$

Let us briefly explain how we can recover v for a given p. We sort a set of portfolio values $\{V_i\}$ in an ascending order, then starting with $k = 1$ keep computing $\hat{p}(k) = \frac{1}{N}\sum_{i=1}^{k}\frac{f(X_i)}{\pi(X_i)}$ (or $\frac{\sum_{i=1}^{k} f(X_i)/\pi(X_i)}{\sum_{i=1}^{N} f(X_i)/\pi(X_i)}$ for π with unknown normalization constant) until $\hat{p}(k) = p$. VAR is given by $v = V_k$. Its error can be computed by order statistics.

When π is picked so that the expression $I_{B_v}(x)\frac{f(x)}{\pi(x)}$ is nearly constant under the probability density π, then the variance will be reduced significantly if not eliminated. This can be achieved when π is similar to the function $I_{B_v}(x)f(x)$, or if π samples more in B_v. Various authors proposed methods for creating π. Glasserman, Heidelberger, and Shahabuddin ([35],[36],[40],[41], henceforth GHS) computed new twisted density with new mean and covariance using analytic as a guide (See Appendix B for more details). This method does not work well if the analytic approximation is not accurate, or if there is more than one local minimum. Glass([32]) proposed non-linear optimization to identify local minima, and construct a Gaussian twisted density based on these local minima. This method works for the cases when the GHS method fails. In the next chapter, we will propose a new importance twist π by extending Glass's approach.

2.2 Control Variate Method for VAR calculation

The control variate method is an another variance reduction technique where the original estimator is replaced by another estimator with smaller variance plus a deterministic formula. The idea is to find a function that is similar to the integrand of the integration in question. The control variate method calls for a function $k(x)$ so that $\int_{R^d} k(x)f(x)dx$ is analytically known and finite. Furthermore, the method calls for the variance of $(I_{B_v}(x) - k(x))$ to be smaller than the variance of the original $I_{B_v}(x)$ under the probability density f. Then the probability p can be expressed as:

$$p = \int_{R^d}(I_{B_v}(x) - k(x))f(x)\,dx + \int_{R^d} k(x)f(x)dx \tag{2.4}$$

and its estimator is given by:

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N}(I_{B_v}(X_i) - k(X_i)) + \int_{R^d} k(x)f(x)dx \tag{2.5}$$

If k is similar to I_{B_v}, then the variance reduction can be achieved. This method is a popular method in pricing (i.i, the Asian options), and is not a good method for rare event simulation in general. Nevertheless, it performs better than the importance sampling method for some linear portfolios with non-extreme quantile cases ($p \geq 0.05$), so we shall discuss how this method works. One candibate for k is $k \equiv I_{\hat{B}_v}$ where $\hat{B}_v = \{\hat{Q} \leq v\}$ and $\hat{Q}$ is the Taylor approximation of V given by:

$$\hat{Q} = \frac{\partial V}{\partial t}\Delta t + (\frac{\partial V}{\partial S})^T \cdot \Delta S + \frac{1}{2}\Delta S^T \cdot (\frac{\partial^2 V}{\partial S^2}) \cdot \Delta S \tag{2.6}$$

where Δt is a horison and ΔS is a vector of changes in underlying risk factor S whose notation is given in section (1.4). The partial derivative expressions for the $\frac{\partial V}{\partial S}$ is a gradient, and $\frac{\partial^2 V}{\partial S^2}$ is a Hessian matrix of V. Since $\hat{Q}$ is the Taylor expansion of V, this choice is a reasonable one. However, this method fails when V is highly non-linear. Juan Cardenas, Emmanuel Fruchard, and others ([9]) showed that for this $\hat{Q}$, $\int_{R^d} k(x)f(x)dx$ can be semi-analytically computed using Fourier transform method (See Appendix A).

Chapter 3

New Importance Sampling Technique via I.I.D. Monte Carlo

3.1 Introduction

In this chapter, we will propose new importance sampling strategies for computing VAR. The natural choice for a good importance sampling twisted density π would be to create one from the original density f with shifted mean(s). If the new mean is chosen so that it samples more near the region of interest, then it would be a successful importance twisted density. The major question is how one can compute or estimate new mean and covariance.

Most importance sampling strategies applied to VAR are based on the main premise of the Large Deviation Theories (Dembo [17]). Twisted density with new mean should sample most data around the new mean so that $V \leq v$ most of the time. Our objective is to find the point(s) so that the density f is maximized on the VAR region $V \leq v$. To find the most likely way to get $V \leq v$, we (following Glass [32]) perform a nonlinear optimization:

$$\max_{x \in B_v} f(x) \tag{3.1}$$

where [1] x is a d-dimensional normal process in R^d, and $B_v = \{x \in R^d : V(x) \leq v\}$. β such that $f(\beta) = \max_{x \in B_v} f(x)$ is called the 'local minimum', or the 'MRP' ('Minimum Rate Point' as called by Bucklew [6]). For some portfolios, there are multiple local minima or MRPs. Glass found, and we also confirm, that multiple local minima do occur and it is important to use them all for the importance

[1]Glass [32] performed maximization on $f(S)$ which is a correlated lognormal density. Instead, we will perform optimization on $f(x)$ as a correlated normal density in R^d.

sampling. We will then introduce a new exponential twisted density in the direction perpendicular to the VAR surface $V = v$ at each local minimum. We will show the effectiveness of this method both experimentally and analytically.

3.2 Non-Linear Optimization for Maximizing Density $f(x)$

Let us describe our optimization technique in more detail. We hope to maximize the density $f(x)$ in a set B_v where the density f is given by:

$$f(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_x|^{1/2}} e^{-\frac{1}{2}x^T \cdot \Sigma_x^{-1} \cdot x} \tag{3.2}$$

where Σ_x is a covariance matrix. This can be done by an optimization technique called the Penalty Method. The penalty method is an iterative routine where a penalty factor a is introduced to perform a global unconstrained minimization as follows:

$$\min_{x \in R^d} \Psi(x, a) \tag{3.3}$$

where Ψ is given by:

$$\Psi(x, a) = -f(x) + a\,(V(x) - v)^2 \tag{3.4}$$

where $a > 0$ is a penalty constant. At each iterative step, the penalty factor a is increased until the the solution to the problem (3.1) is found. We shall not describe the actual optimization algorithm here since the theories and practices are explained in many optimization text books ([3], [59], [71]). We have used 'fminsearch' function from Matlab (v12.1) to perform global optimization for the above penalty method.

Some portfolio have multiple local minima. In order to counter this problem, one can make systematic search by dividing the space R^d into several blocks (for example, each quadrant). Then a set A of candidates for the initial point is picked from each of these blocks. This can be done heuristically, and the above optimization can be performed for each of the element of the set A. The optimization is computationally very inexpensive, and thus running this process several times is not an issue compared with the actual Monte Carlo runs. Furthermore, once could pick an initial guess by resorting to some analytic method. As a comparison, we have used the GHS method (Appendix B, [41]) to pick a starting initial point. This method picks a point in R^d near the region $V = v$ if V does not have multiple local minima.

3.3 Importance Twisted Density for Computing VAR Probability

Suppose we have m local minima, say $\beta = \{\beta_1, \beta_2, ..., \beta_d\}$. Then for each of these local minima, we construct the following Gaussian twist:

$$\pi(x) = \sum_{i=1}^{m} \frac{\alpha_i}{(2\pi)^{d/2}|\Sigma_x|^{1/2}} e^{-\frac{1}{2}(x-\beta_i)^T \cdot \Sigma_x^{-1} \cdot (x-\beta_i)} \tag{3.5}$$

where $x \in R^d$, and α_i is a weighting factor such that $\sum_{i=1}^{m} \alpha_i = 1$. Each weighting α_i should be selected according to the relative importance of each local minima. Therefore, we have used the following weighting using the original density $f(x)$ as a guide:

$$\alpha_i = \frac{f(\beta_i)}{\sum_{k=1}^{m} f(\beta_k)} \tag{3.6}$$

The intuitive idea of this weighting is to sample more where the density is large. One could also assign the equal weight $\frac{1}{m}$ to each of α_i. We confirmed that this Gaussian twist works for all portfolios (chapter 4) if all important local minima are sampled using the density (3.5). This is guaranteed by the Large Deviation Theory as well. In the next section, we will propose a better twisted density function that will boost the performance even further.

3.4 Exponential Density as the New Importance Twist

Let us assume that all local minima are identified β_i , $i = 1, 2, ..., m$ via non-linear optimization as described in the previous section. Then we propose an exponential twisted density in the direction perpendicular to the VAR surface $V = v$ at each local minimum β_i, and Gaussian densities in other directions. Let us call this direction x_{β_i} which is given by $\Sigma_x^{-1} \cdot \beta_i$ (The derivation will be given shortly). Let us define the unit vector for x_{β_i} as u_1, and call the coordinate as x_1. We can construct new orthogonal coordinates starting with x_1 by performing Gram-Schmidt orthogonalization process. Let us define the corresponding unit vector for each $x_{k,k=2,3,...,d}$ as u_k. At each local minimum β_i, the VAR surface $V = v$, and the density surface $f = C$ (constant) is tangent to each other. If the VAR region B_v is convex, then we can place the entire VAR region B_v inside a half space H so that its boundary ∂H called 'Hyperplane' is tangent to both B_v, and the density surface f at β_i where H is defined as $H = \{x \in R^d : x_{\beta_i} \cdot (x - \beta_i) \geq 0\}$. Then an exponential density can be constructed starting from this hyperplane ∂H (that is to start from $|\beta_i^T \cdot u_1|$

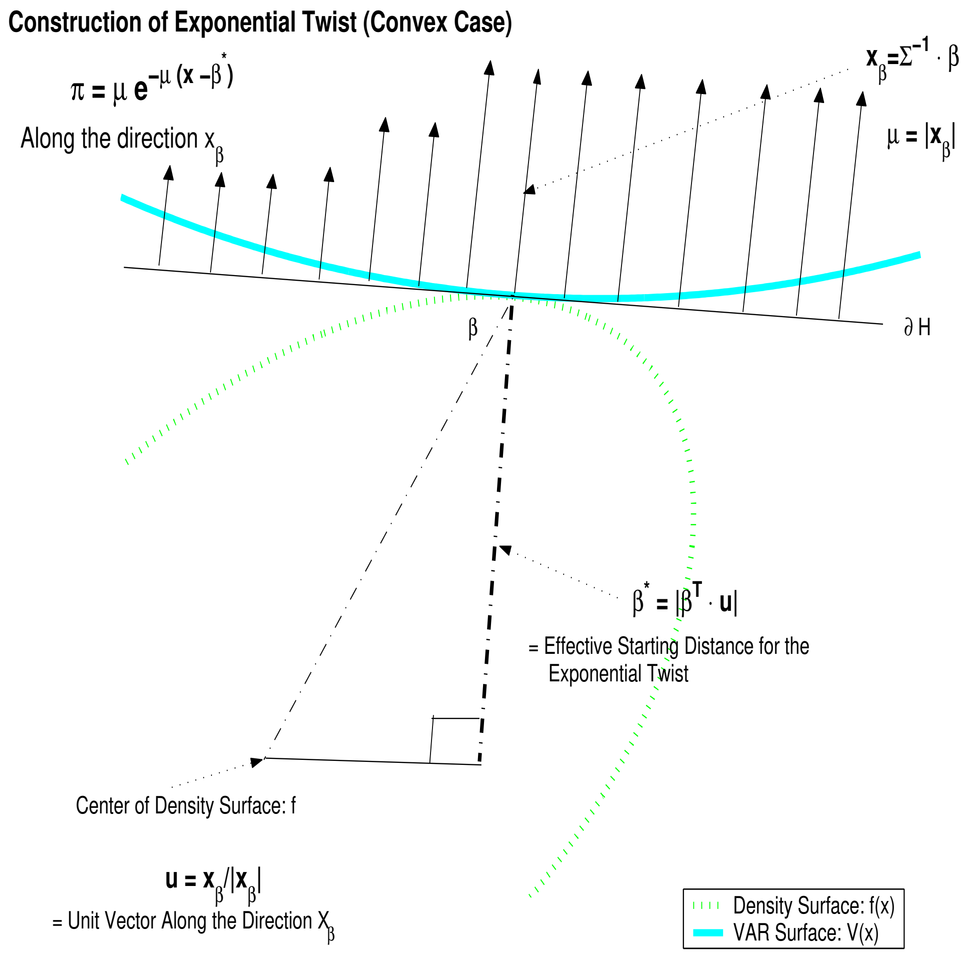

Figure 3.1: When the VAR region B_v is convex, and is contained in the half space H. Then the exponential twist is created along the direction perpendicular to the VAR surface (or Density surface f) at β. In this case, β^* is a fixed number computed by $|\beta^T \cdot u|$ where u is a unit vector along the perpendicular direction x_β. The Gaussian twists are created for all other orthogonal directions.

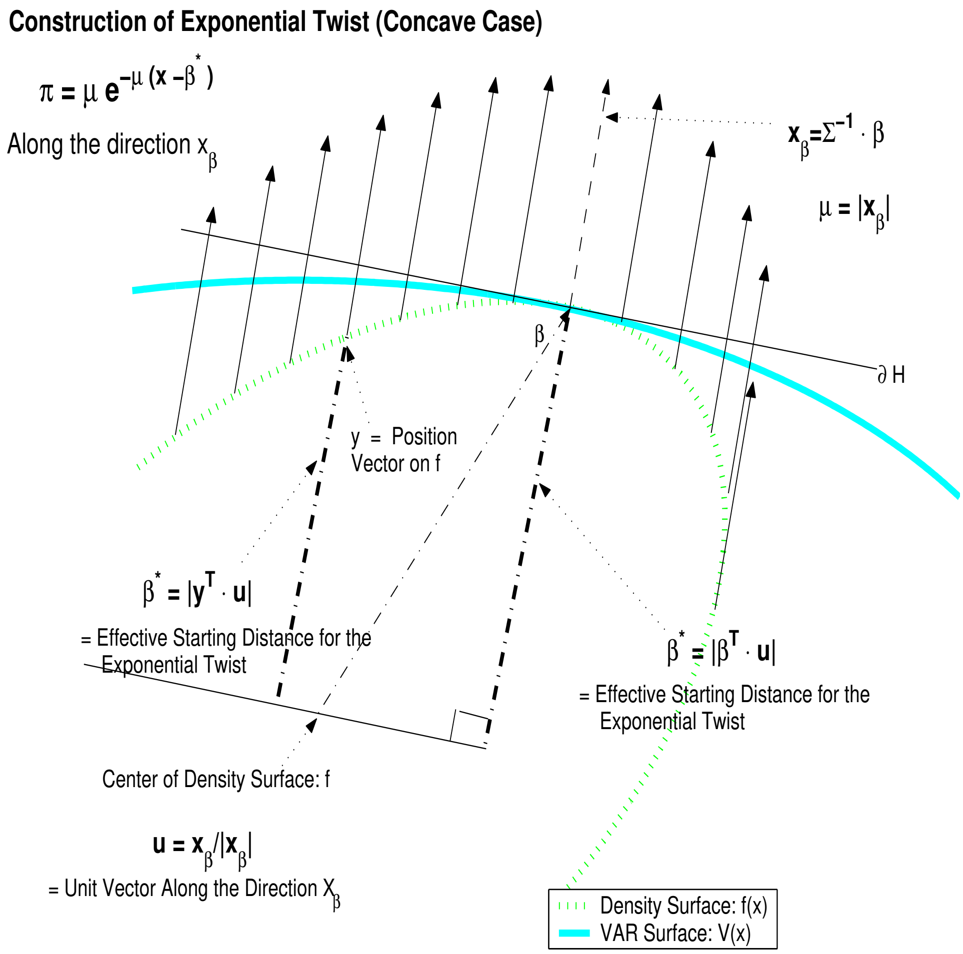

Figure 3.2: When the VAR region B_v is concave, then the exponential twist is still created along the direction perpendicular to the VAR surface (or Density surface f) at β. However this time, β^* must be changed at every point y on the density surface f. β^* is now given by $|y^T \cdot u|$. Likewise the convex case, the Gaussian twists are created for all other orthogonal directions.

which is the projected length of β_i along the direction of u_1, See Figure 3.1). If B_v is concave and some part of B_v comes closer to f bypassing the hyperplane H, then we need to construct an exponential density starting from the boundary of f (which is the ellipse $x^T \cdot \Sigma_x^{-1} \cdot x^T = \beta_i^T \cdot \beta_i$) to make sure that all VAR region is covered (See Figure 3.2). Thus, motivated by the Large Deviation Theory and the Laplace method (Appendix C), we propose the following density along x_1:

$$\pi_e^i(x_1) = \mu e^{-\mu(x_1 - \beta_i^*)} \tag{3.7}$$

where [1] $\mu = |x_{\beta_i}|$, $x_1 \geq \beta_i^*$, and β_i^* is given by:

$$|\beta_i^*| = \begin{cases} |\beta_i^T \cdot u_1| & \text{if } B_v \in H \text{ (i.e.,} B_v \text{ is convex)} \\ |y^T \cdot u_1| & \text{otherwise (i.e.,} B_v \text{ is concave).} \end{cases} \tag{3.8}$$

where y is the point on the ellipse: $y^T \cdot \Sigma_x^{-1} \cdot y^T = \beta_i^T \cdot \beta_i$. Rather than finding y, once can compute $|y^T \cdot u_1|$ directly by $|y^T \cdot u_1| = b$. where $b = \sqrt{\beta_i^T \cdot \beta_i - \bar{x}^T \cdot \Sigma_x^{-1} \cdot \bar{x}^T}$, and $\bar{x} = (0, x_2, x_3, ..., x_d)$. When b becomes imaginary, then set $b = 0$. For all other $d-1$ directions $x_2, x_3, ..., x_d$, we propose the following Gaussian densitiy:

$$\pi_g^i(x_k) = \frac{1}{\sqrt{2\pi}\lambda_k} e^{\frac{-x_k^2}{2\lambda_{Min}^2}} \tag{3.9}$$

where λ_{Min} is the smallest eigenvalue of the correlation matrix Σ_x. The idea is that in the most important direction u_1, we propose the exponential density, and use the Gaussian densities for all other orthogonal directions. Thus, the final target density is the expressed as:

$$\pi^i(x) = \prod_{k=2}^{d} \pi_g^i(x_k)\, \pi_e^i(x_1) \tag{3.10}$$

If B_v is convex, then the order of sampling x_k does not matter. If B_v is concave, then $\prod_{k=2}^{d} \pi_g^i(x_k)$ must be sampled first in order to find out the starting boundary β_i^* for $\pi_e^i(x_1)$. Finding out the curvature (whether B_v is concave or convex at β) is quite simple. Take any small arbitrary $\varepsilon > 0$, and any arbitrary unit direction $u_j \neq u_1$. If $V(\beta_i + \varepsilon u_j) > v$, then it is convex. Otherwise, it is concave. This statement is locally true since β_i is assumed to be a local minimum. One can perform the same construction for all other local minima β_i to get the following effective final target density for the multiple local minima case:

$$\pi(x) = \sum_{i=1}^{m} \alpha_i \prod_{k=2}^{d} \pi_g^i(x_k)\, \pi_e^i(x_1) \tag{3.11}$$

[1] $||$ is the Euclidean norm in R^d

Note that x for each local minima β_i has its own coordinate system. Therefore, in order to construct the likelihood function $\frac{f(x)}{\pi(x)}$, we compute the effective vector in the original coordinate system as:

$$x = \sum_{i=1}^{d} u_i x_i \tag{3.12}$$

and plug this into $f(x)$ to compute $\frac{f(x)}{\pi(x)}$. As long as we cover all local minima, we do not have to worry about concave VAR regions (if any) crossing the density surface $f = C$ other than touching β_i. If this happens then any points where V surface crosses $f = C$ will be more optimal than β_i (implying the existence of much smaller $f = C$ curve tangential to these points) contradicting the fact that β_i is the local minimum. Or these points will be one of other local minima being found via optimization. Therefore the weighted density (3.11) should work well.

We shall briefly discuss how we derived the expression for x_{β_i} as $\Sigma_x^{-1} \cdot \beta_i$. The gradient $\nabla V(\beta_i)$ gives the perpendicular direction at β_i. Since both the density surface $f = C$ and the VAR surface $V = v$ is tangent to each other at each local minima β_i, the gradient $\nabla f(\beta_i)$ can be used as well. So computing $\nabla f(\beta_i)$ gives:

$$\begin{aligned}
\nabla f(\beta_i) &= f(\beta_i)\nabla(-\frac{1}{2}x^T \cdot \Sigma_x^{-1} \cdot x)|_{x=\beta_i} \\
&= -f(\beta_i)\Sigma_x^{-1} \cdot \beta_i
\end{aligned}$$

Thus the direction is along the constant multiple of the vector $\nabla f(\beta_i)$. The alternative derivation is based on the Large Deviation Theory. As will be shown later, at each local minima (MRP) β_i, x_{β_i} must satisfy $\nabla\Lambda(\beta_i) = x_{\beta_i}$ where Λ is a logmoment generating function given by $\Lambda(\beta_i) = \log E[e^{\beta_i \cdot x}]$. Since $E[e^{\beta_i \cdot x}]$ is computed as:

$$\begin{aligned}
E[e^{\beta_i \cdot x}] &= \int_{R_d} \frac{1}{(2\pi)^{d/2}|\Sigma_x|^{1/2}} e^{-\frac{1}{2}x^T \cdot \Sigma_x^{-1} \cdot x + \beta_i \cdot x} dx \\
&= e^{\frac{1}{2}\beta_i^T \cdot \Sigma_x^{-1} \cdot \beta_i}
\end{aligned}$$

$\Lambda(\beta_i) = \log E[e^{\beta_i \cdot x}] = \frac{1}{2}\beta_i^T \cdot \Sigma_x^{-1} \cdot \beta_i$. Therefore, $x_{\beta_i} = \nabla\Lambda(\beta_i) = \Sigma_x^{-1} \cdot \beta_i$. It turn out experimentally, that this factor is confirmed to be an optimal one. Too big or too small x_{β_i} will result in sampling too much to too little in the far region of B_v. Note that when f is an uncorrelated Gaussian density, then $x_{\beta_i} = \Sigma_x^{-1} \cdot \beta_i = \beta_i$. This makes sense since vector from the origin to any points on the sphere is perpendicular to the sphere itself.

Next, we shall explain the effectiveness of our exponential twist over the Gaussian twist. The Exponential twisting has several advantages. The first advantage is that the tail of the distribution for f behaves like an exponential distribution. Let us take a simple one dimensional example where our objective is to estimate $Pr(X > v)$ where $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ for large $v = 2.326$ which corresponds to $p = 1$. The Gaussian twist will be given by $\pi = \frac{1}{\sqrt{2\pi}}e^{-(x-v)^2/2}$, and our exponential twist is $\pi = ve^{-(x-v)v}$. The graph of these densities is shown in the figure (3.3). The figure shows that f falls off much faster than the shifted Gaussian when x goes beyond v. Thus, most of the Gaussian twist samples too far from v to contribute much to the over all p. On the other hand, the exponential distribution approximate the original f very well.

Let us analyze more quantitatively. By following the same reasoning (but much simpler) as done in the Appendix C, we can compute p asymptotically as $p = e^{-\frac{v^2}{2}}\frac{1}{\sqrt{2\pi}v}[1-\frac{1}{v^2}+\frac{3}{v^4}]+o(\frac{1}{v^7})$. The second moment M_g^2 for the Gaussian twist can be computed similarly as $M_g^2 = e^{-v^2}\frac{1}{\sqrt{2\pi}2v}[1-\frac{1}{(2v)^2}+\frac{3}{(2v)^4}]+o(\frac{1}{v^7})$. Likewise, the second moment M_e^2 for the exponential twist is given by $M_e^2 = e^{-v^2}\frac{1}{2\pi v^2}[1-\frac{2}{v^2}+\frac{12}{v^4}]+o(\frac{1}{v^7})$. So the variance for the Gaussian twist is computed as $M_g^2-p^2 = \frac{e^{-v^2}}{\sqrt{2\pi}v}[\frac{1}{2}-\frac{1}{\sqrt{2\pi}v}+o(\frac{1}{v^2})]$, and the variance for the exponential twist is computed as $M_e^2-p^2 = \frac{5e^{-v^2}}{2\pi v^6}[1+o(\frac{1}{v^8})]$. We can notice that the variance ratio for the exponential twist drops significantly because of the cancellation of many leading terms of v resulting with $\frac{1}{v^6}$ being the largest leading term. On the other hand, the leading term for the Gaussian twist is $\frac{1}{v}$. Computing the variance ratio of Gaussian twist over the exponential twist is given by $\frac{M_g^2-p^2}{M_e^2-p^2} = \frac{\sqrt{2\pi}v^5}{5}[\frac{1}{2}-\frac{1}{\sqrt{2\pi}v}+o(\frac{1}{v^2})]$. Thus for our model problem, $v = 2.326$, so $\frac{M_g^2-p^2}{M_e^2-p^2} \sim 11.21$ which is consistent with most of our experimental results (Appendix E). The variance ratios for the Gaussian twist is somewhere between 10 and 50 whereas the variance ratios for the exponential twist is generally around 3 digits (100 to 500). Asymptotically, the exponential twist outperform the Gaussian counterpart by order of v^5 as v becomes large since $\frac{M_g^2-p^2}{M_e^2-p^2} \sim \frac{\sqrt{2\pi}v^5}{5}$. Therefore, the exponential twist is the best candidate for the rare event simulation.

The second advantage of the exponential density is that it is at least efficient by almost a factor of 2. This is because of the fact that the exponential density samples one side of the VAR boundary $V = v$, i.e., $V \leq v$ most of the time. On the other hand, the Gaussian twist samples both regions $V \leq v$ and $V \geq v$ although it is peaked at $V = v$. This conceptually results in half of the samples being wasted. Furthermore, the exponential twist avoids sampling near the origin which typically gives huge likelihood ratios resulting in larger variance. We shall show an example where the importance sampling with the exponential twist outperforms the Gaussian twist for multiple local minima case. Take for example a portfolio V consists of butterflies in 2 dimension (one long call struck at 90, one long call struck at 110, and 2 short calls struck at 100). It is well known that this portfolio have two ways to lose money in each direction. Let us assume that a trader hold butterfly trades in two different underliers each uncorrelated. This will results in four local minima. Let us take a correlated case (say, correlation = 0.8). Then the optimization finds (with $p = 0.01$) two local minima as shown in the figures (3.4), and (3.5). One can see that the underlying density looks like an ellipse (no longer a circle), and aligned along

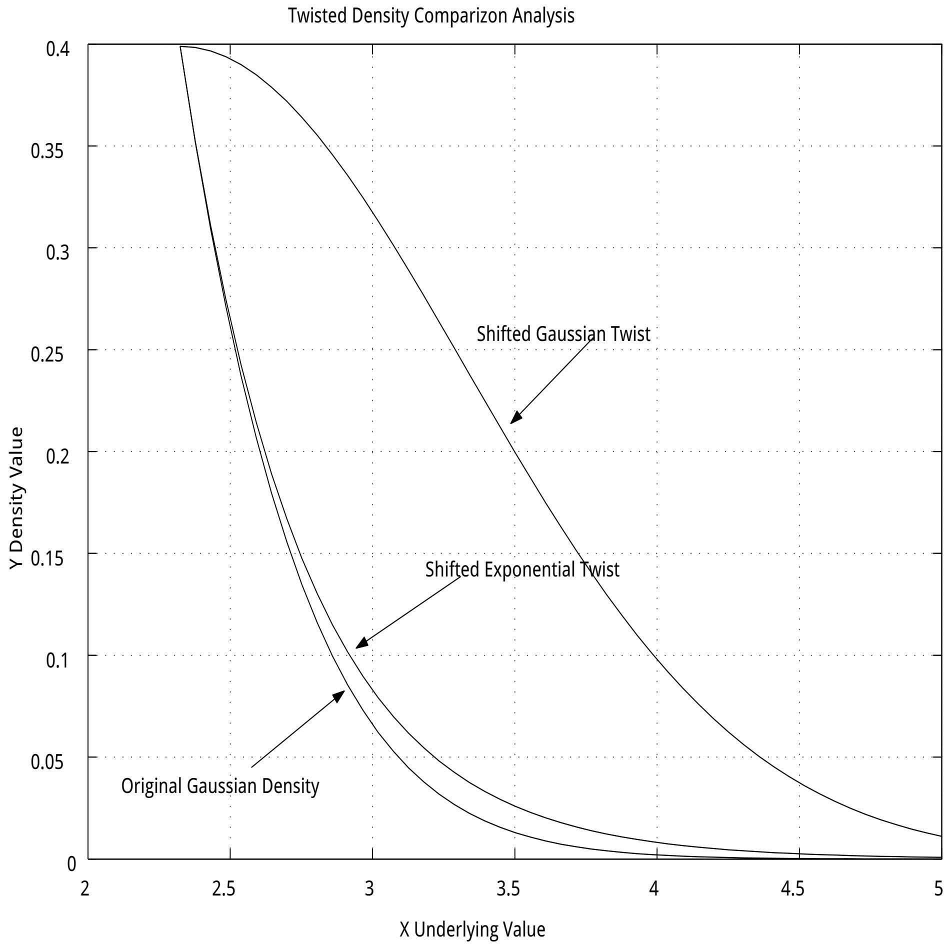

Figure 3.3: Density Plot: The Original Density $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, the Gaussian twisted density with shifted mean (scaled) $\frac{1}{\sqrt{2\pi}}e^{-(x-v)^2/2}$, and the exponential density with shifted mean (also scaled) $ve^{-(x-v)v}$ near the 99th quantile (v=2.326) are plotted. Note that the exponential twist seems to better approximate the behavior of the original function than the Gaussian counterpart. The Gaussian twist samples too far outside the true region of interest. Therefore, it results in a loss of computational time.

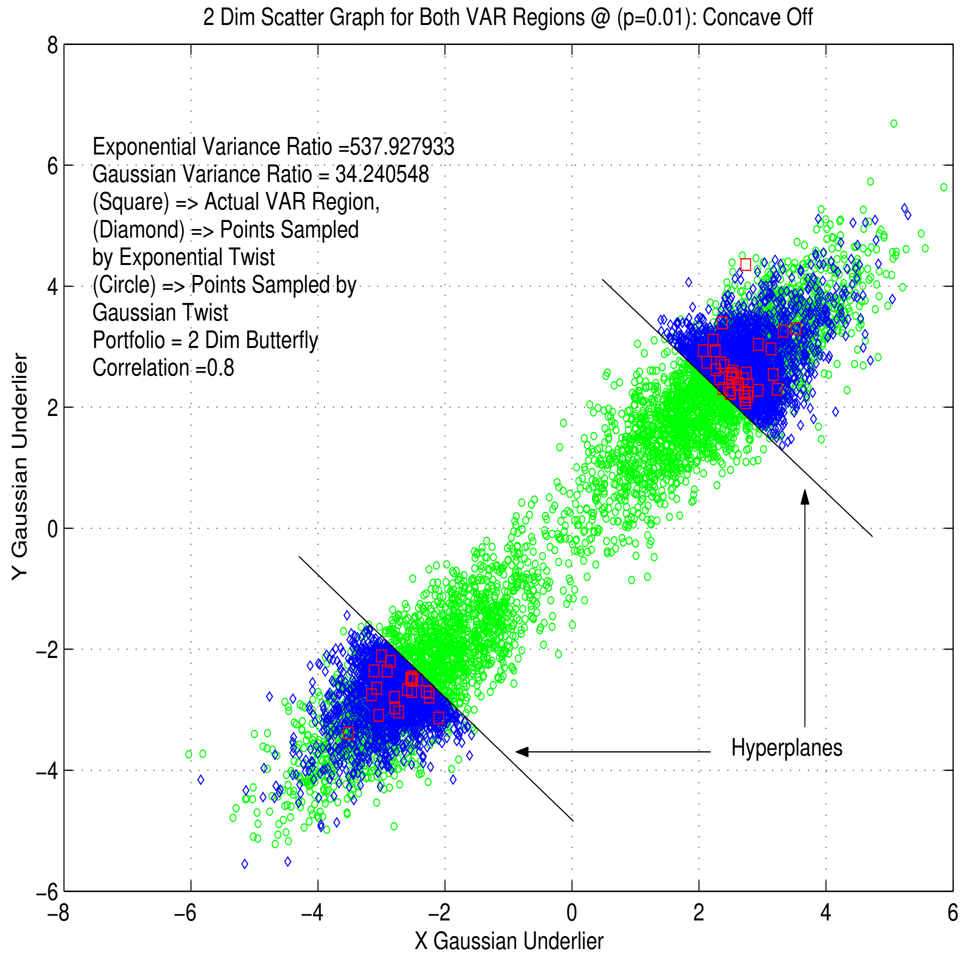

Figure 3.4: Scatter plot for the simulated V with correlation = 0.8 in 2 dimension (one long call struck at 90, one long call struck at 110, and 2 short calls struck at 100). Here, we have butterfly position in each dimension. For this case, there are 2 local minima as can be seen from the graph. Corresponding local minima are $\beta_1 = [2.33203E{+}00, 2.33196E{+}00]$, and $\beta_2 = [-2.42727E{+}00, -2.42727E{+}00]$. Since the VAR region is convex, one construct an exponential twist starting from the hyperplane ∂H. It is clear from the graph, more than the half of the Gaussian twist samples fall out side the VAR region. The Gaussian twist gives the variance ratio of about 34. On the other hand, the effectiveness of the exponential twist can be seen both visually, and numerically. This twist samples heavily in the actual VAR region, and the variance ratio for this portfolio is over 530. Furthermore, the exponential twist avoids sampling near the origin which typically gives huge likelihood ratios resulting in larger variance.

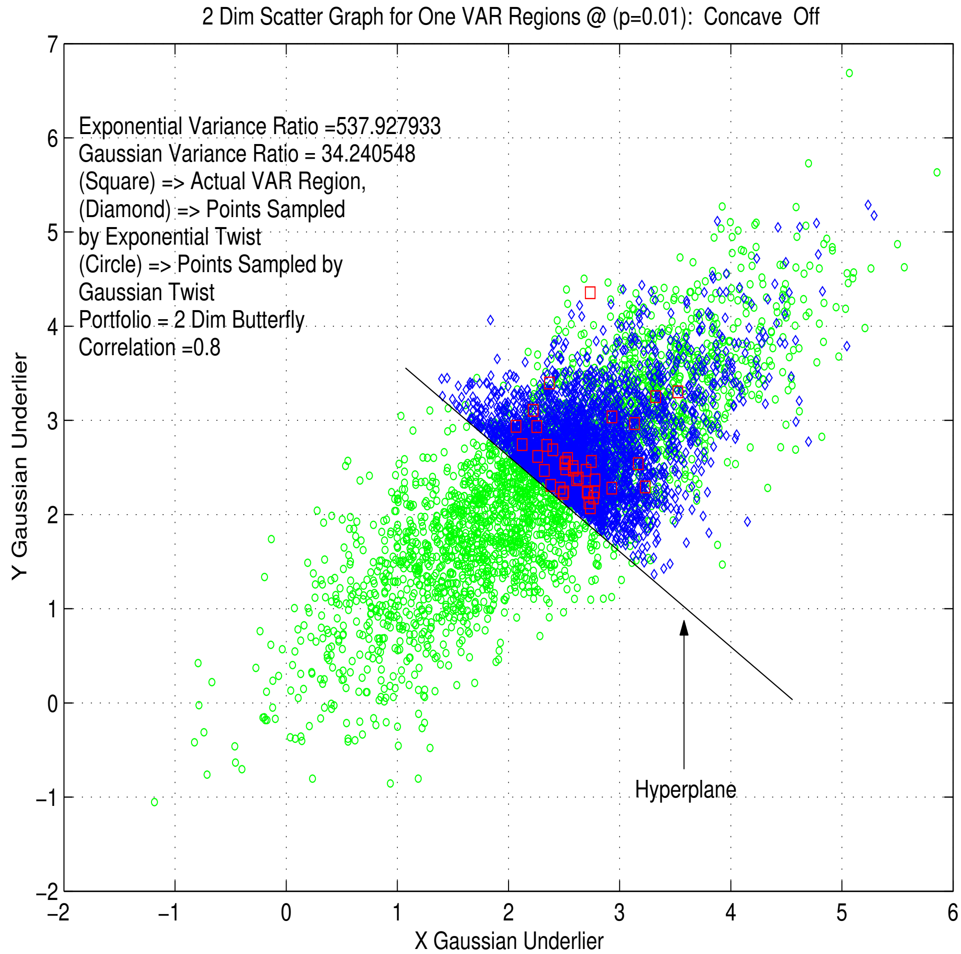

Figure 3.5: (One Local Minimum VAR Region Magnified) The Scatter plot for the simulated V with correlation $= 0.8$ in 2 dimension (one long call struck at 90, one long call struck at 110, and 2 short calls struck at 100). Here, we have butterfly position in each dimension. Since the VAR region is convex, one construct an exponential twist starting from the hyperplane ∂H. It is clear from the graph, more than the half of the Gaussian twist samples fall out side the VAR region. The Gaussian twist gives the variance ratio of about 34. On the other hand, the effectiveness of the exponential twist can be seen both visually, and numerically. This twist samples heavily in the actual VAR region, and the variance ratio for this portfolio is over 530.

the line $x = y$. This makes sense since high correlation has the effect of collapsing VAR region together. In this example, B_v is convex, so the exponential twist is formed from the hyperplane ∂H. As can be seen from the figure, two exponential twists are used to construct the effective twist π. It is visually clear from the figure (3.4) and (3.5) that the Gaussian twist wastes too many samples (over half) although the variance ratio is still good (about 30). The exponential twist, on the other hand, samples almost within the actual VAR region, and the variance ratio is about 530. Therefore, this example has shown that the exponential twist outperforms the Gaussian twist even for the multiple local minima case. Furthermore, this result confirms that it is important to use all local minima to create the exponential twist.

3.5 Large Deviation Analysis

The tail behavior of the distribution can be best described by the Large Deviation Theory (Henceforth LDT). From the LDT context, the VAR probability for the rare event case can be expressed as :

$$p = P_n = P(\frac{1}{n}\sum_{i=1}^{n} V_i \leq v) = \int_{R^d} I_{B_v}(x) f_n(x) dx \tag{3.13}$$

For the importance sampling, we have the following representation for the second moment: $M_n^2(\pi_n) = \int_{R^d}[I_{B_v}(x)\frac{f_n(x)}{\pi_n(x)}]^2 \pi_n(x)dx$. By applying the Jensens inequality, we have:

$$M_n^2(\pi_n) \quad \geq \quad [\int_{R^d} [I_{B_v}(x)\frac{f_n(x)}{\pi_n(x)}]\pi_n(x)dx]^2 = P_n^2 \tag{3.14}$$

So the variance $= M_n^2(f_n) - P_n^2$ can be minimized by letting $I_{B_v}(x)\frac{f_n(x)}{\pi_n(x)} = P_n$.

For each $\beta \in R^d$, let us define the following rate function $\Lambda_n(\beta) = \frac{1}{n}\log(E[e^{n\beta\cdot X_n}])$ where the expectation is taken under the density f_n. The LDT calls for defining asymptotic log-moment generating function as : $\Lambda(\beta) = \lim_{n\to\infty} \Lambda_n(\beta)$. In order to avoid technical details, let us assume that the above limit exist which is the case when f is correlated Gaussian density (Bucklew [6]). Let us define the following Large Deviation rate function called the Legendre-Frenchel transform of $\Lambda(\beta)$ as: $\Lambda^*(x) = \sup_{\beta\in R^d}(\beta \cdot x - \Lambda(\beta))$. and the essential domain of $\Lambda(x)$ as the set $D = \{\beta \in R^d : \Lambda(\beta) < \infty\}$. Furthermore, let us define the Cramer transform of a Borel set B_v as: $\Lambda^*(B_v) = \inf_{x\in B_v} \Lambda^*(x)$. Since $\Lambda(\beta)$ is strictly convex and differentiable (a logmoment function under the Gaussian density) on all of R^d, the LDT guarantees the followings ([17], [6]): (1) D is a convex set and non-empty, (2) $\Lambda^*(y)$ is strictly convex on the interior of D (denoted D^o), and lower semicontinuous (3) For each $\beta \in D^o$, there exists an unique $x_\beta \in R^d$

such that $\nabla\Lambda(x_\beta) = \beta$, and $\Lambda^*(\beta) = x_\beta \cdot \beta - \Lambda(x_\beta)$. Let us assume that the closure of B_v (denoted B_v^o) equals to the interior of B_v, and $B_v^o \cap D^o \neq \emptyset$. Then the LDT further guarantees the existence of the following limit:

$$\lim_{n\to\infty} \frac{1}{n} \log(P_n) = -\Lambda^*(B_v) \tag{3.15}$$

Bucklew ([6]) defined that a sequence π_n of simulation distribution is asymptotically efficient if

$$\lim_{n\to\infty} \frac{1}{n} \log(M_n^2(\pi_n)) = -2\Lambda^*(B_v) \tag{3.16}$$

If π_n is asymptotically efficient, then $P_n \sim e^{-n\Lambda^*(B_v)}$, and $M_n^2(\pi_n) \sim e^{-2n\Lambda^*(B_v)}$. Therefore $P_n^2 \approx M_n^2(\pi_n)$, and the variance will be effectively cancelled to zero. On the other hand, if π_n is not asymptotically efficient, then the relative error grows exponentially since $\frac{\sqrt{M_n^2(\pi_n) - P_n^2}}{P_n} \sim Const \cdot e^{n\Lambda^*(B_v)} \to \infty$. Thus our objective is to find a twist which is asymptotically efficient.

According to Bucklew ([6]), β is called a minimum rate point (or 'MRP') of the set B_v if $\beta \in \partial B_v$, and $\Lambda^*(\beta) = \Lambda^*(B_v) \equiv \inf_{y \in B_v} \Lambda^*(y)$. Furthermore, β is defined as a dominant point of the set B_v if β is an unique point such that: (1) $\beta \in \partial B_v$, (2) there exists a unique x_β such that $\nabla\Lambda^*(x_\beta) = \beta$, and (3) $B_v \subset H(\beta) = \{x : x_\beta \cdot (x - \beta) \ \geq 0\}$ where H is called a half space, and ∂H is called the 'hyperplane' as introduced in the previous section when we set up exponential twist off this ∂H. From the convex function theory, the hyperplane $\partial H(\beta) = \{x : x_\beta \cdot (x - \beta) \ = 0\}$ is tangent to the rate function level set $\{x : \Lambda^*(x) = \Lambda^*(\beta)\}$ at the point β, and $\nabla\Lambda^*(x_\beta) = \beta$. Thus if β is a dominant point it is a unique minimum rate point. Therefore, if B_v is convex, and $B_v \cap D^o \neq \emptyset$, then set B_v is covered by the half space H and the dominant point β exists. For this dominant point β, the second moment can be bounded as:

$$M_n^2(\pi_n) \leq \int e^{2n\ x_\beta \cdot (x-\beta)} [\frac{f_n(x)}{\pi_n(x)}]^2 \pi_n dx \equiv \bar{M}_n^2(\pi_n) \tag{3.17}$$

This is so since $I_{B_v} \leq 2n\ x_\beta \cdot (x - \beta)$ for the B_v side of the hyperplane H. Since $\Lambda_n(\beta) = \frac{1}{n}\log(E[e^{n\beta \cdot X_n}])$, and $\int e^{n\ x_\beta \cdot x} f_n(x)dx = E[e^{n x_\beta \cdot X_n}] \equiv e^{n\ \Lambda_n(x_\beta)}$, the above second moment can be expressed as:

$$\begin{aligned} \bar{M}_n^2(\pi_n) &\geq [\int e^{n\ x_\beta \cdot (x-\beta)} [\frac{f_n(x)}{\pi_n(x)}] \pi_n dx]^2 \\ &= [\int e^{n\ x_\beta \cdot (x-\beta)} f_n(x) dx]^2 \\ &= [e^{n\Lambda_n(\beta)} e^{-n\ x_\beta \cdot \beta}]^2 = [e^{n\ (\Lambda_n(x_\beta) - x_\beta \cdot \beta)}]^2 \end{aligned} \tag{3.18}$$

Thus the minimun is obtained by letting: $e^{n\ x_\beta \cdot (x-\beta)} [\frac{f_n(x)}{\pi_n(x)}] = e^{n\ (\Lambda_n(x_\beta) - x_\beta \cdot \beta)}$ So the optimal twist is obtained by: $\pi_n(x) = f_n(x) e^{n\ (x_\beta \cdot x - \Lambda_n(x_\beta))}$. Therefore the general twisting is hinted as: $\pi(x) = f(x)e^{x_\beta \cdot x - \Lambda(x_\beta)}$. For the multivariate Gaussian, $E(e^{x_\beta \cdot x}) = \int_{R^d} \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\sum_{i=1}^d x_i^2} e^{x_\beta \cdot x} dx = e^{\frac{1}{2}x_\beta^T \cdot x_\beta}$ Thus, the twisting becomes: $\pi(x) = f(x)e^{\beta \cdot x - \frac{1}{2}x_\beta^T \cdot x_\beta}$. which amounts to shifting the original density by the dominant point β since $x_\beta = \Sigma_{-1} \cdot \beta$.

Next we shall show that even if there are multiple local minima, the asymptotical efficiency can still be achieved if all local minima are taken into consideration. Let us assume $\Lambda^*(B_v) = \inf_{x \in B_v} \Lambda^*(x) > 0$, and consider a series of distribution of the form: $\pi_n^* = f_n(x) \sum_{i=1}^m \alpha_i e^{n\ (x_{\beta_i} \cdot x - \Lambda_n(x_{\beta_i}))}$ which is a convex combination of twisted densities covering all minimum rate points where $\sum \alpha_i = 1$, and $\bar{\beta} = (\beta_1, \beta_2, ..., \beta_m) \in D^o$. We shall show that if $\bar{B}_v$ (the closure) $\subset \cup_{i=1}^m H(\beta_i)$, and $\Lambda^*(x_\beta) \geq \Lambda^*(B_v)$ for each $i = 1, 2, ..., m$, then π_n^* is asymptotically efficient. (Sufficient Condition). Let us compute $M_n^2(\pi_n)$ as:

$$
\begin{aligned}
M_n^2(\pi_n) &= \int_{I_{B_v}} [\frac{f_n(x)}{\pi_n(x)}]^2 \pi_n(x) dx \leq \sum_{i=1}^m \int_{H(\beta_i)} [\frac{f_n(x)}{\pi_n(x)}]^2 \pi_n(x) dx \\
&= \sum_{i=1}^m \int_{H(\beta_i)} [\sum_{i=1}^m \alpha_i e^{n\ (x_{\beta_i} \cdot x - \Lambda_n(x_{\beta_i}))}]^{-2} \pi_n(x) dx \\
&\leq \sum_{i=1}^m \int_{H(\beta_i)} [\alpha_i e^{n\ (x_{\beta_i} \cdot x - \Lambda_n(x_{\beta_i}))}]^{-2} \pi_n(x) dx \\
&= \sum_{i=1}^m \alpha_i^{-2} \int_{H(\beta_i)} e^{-2\ n\ (x_{\beta_i} \cdot x - \Lambda_n(x_{\beta_i}))} \pi_n(x) dx \\
&= \sum_{i=1}^m \alpha_i^{-2} e^{-2\ n\ (x_{\beta_i} \cdot \beta_i - \Lambda_n(x_{\beta_i}))} \int_{H(\beta_i)} e^{-2\ n\ x_{\beta_i} \cdot (x - \beta_i)} \pi_n(x) dx \\
&\leq \sum_{i=1}^m \alpha_i^{-2} e^{-2\ n\ (x_{\beta_i} \cdot \beta_i - \Lambda_n(x_{\beta_i}))} \qquad (3.19)
\end{aligned}
$$

since $[\sum_{i=1}^m \alpha_i e^{n\ (x_{\beta_i} \cdot x - \Lambda_n(x_{\beta_i}))}]^{-2} \leq [\alpha_i e^{n\ (x_{\beta_i} \cdot x - \Lambda_n(x_{\beta_i}))}]^{-2}$, and $x_{\beta_i} \cdot (x - \beta_i) \geq 0$ for $\beta_i \in H(\beta_i)$. Since $\forall i$, $\lim_{n \to \infty} (x_{\beta_i} \cdot \beta_i - \Lambda_n(x_{\beta_i})) \to (x_{\beta_i} \cdot \beta_i - \Lambda(x_{\beta_i})) = \Lambda^*(\beta_i)$, we have: $\lim\sup_{n \to \infty} \frac{1}{n} log(M_n^2(\pi_n)) \leq -2 \min_{i=1,2,...,m} \Lambda^*(\beta_i) = -2\Lambda^*(B_v)$. This is enough to prove the sufficient condition since the LDT guarantees that $\lim\inf_{n \to \infty} \frac{1}{n} log(M_n^2(\pi_n)) \geq -2\Lambda^*(B_v)$.

Let us assume that all MRPs of B_v are in D^o. Then we shall show that π_n^* is asymptotically efficient only if $\bar{\beta}$ contains all minimum rate points of B_v (Necessay Condition) Take $\xi \in B_v^o \cap D^o$, $B_\varepsilon(\xi) = (x : \|x - \xi\| < \varepsilon)$. Since $\xi \in B_v^o \cap D^o$, we can find $\varepsilon : B_\varepsilon(\xi) \subset B_v^o \cap D^o$. Define a twist centered around for ξ as : $\pi_n^\xi(x) = f_n(x)e^{n\ (x_\beta \cdot x - \Lambda(x_\beta))}$. Then

$$
M_n^2(\pi_n) = \int_{I_{B_v}} [\frac{f_n(x)}{\pi_n(x)}]^2 \pi_n(x) dx = \int_{I_{B_v}} \frac{f_n(x)}{\pi_n(x)} f_n(x) dx = \int_{I_{B_v}} \frac{f_n(x)}{\pi_n(x)} \frac{f_n(x)}{\pi_n^\xi(x)} \pi_n^\xi(x) dx
$$

$$\begin{aligned} &\geq \int_{B_\varepsilon(\xi)} [\sum_{i=1}^m \alpha_i [e^{n\ (x_{\beta_i}\cdot x - \Lambda_n(x_{\beta_i}))}]]^{-1} e^{-n\ (x_\xi \cdot x - \Lambda_n(x_\xi)} \pi_n^\xi(x)dx \\ &= e^{-n\ (x_\xi\cdot\xi - \Lambda_n(x_\xi))} \int_{B_\varepsilon(\xi)} [\sum_{i=1}^m \alpha_i [e^{n\ (x_{\beta_i}\cdot x - \Lambda_n(\beta_i))}]]^{-1} e^{-n\ x_\xi\cdot(x-\xi)} \pi_n^\xi(x)dx \end{aligned}$$

For all $x \in B_\varepsilon(\xi)$, we have $x_\xi \cdot (x-\xi) \leq \varepsilon \|x_\xi\|$. Also, $x_{\beta_i}\cdot x = x_{\beta_i}\cdot(x-\xi) + x_{\beta_i}\cdot\xi \leq \varepsilon\|x_{\beta_i}\| + x_{\beta_i}\cdot\xi$. Therefore, we have:

$$M_n^2(\pi_n) \quad \geq \quad e^{-n\ (x_\xi\cdot\xi - \Lambda_n(x_\xi))}[\sum_{i=1}^m \alpha_i [e^{n\ (\varepsilon\|x_{\beta_i}\| + x_{\beta_i}\cdot\xi - \Lambda_n(x_{\beta_i}))}]^{-1} e^{-n\ \varepsilon\|x_\xi\|}\pi_n^\xi(B_\varepsilon(\xi))$$

$\pi_n^\xi(B_\varepsilon(\xi))$ is a probability of $B_\varepsilon(\xi)$ (whose center is ξ) under the shifted twisted density π_n^ξ which is also centered at ξ. Thus, $\pi_n^\xi(B_\varepsilon(\xi)) \to 1$, and we have:

$$\begin{aligned} \lim\inf_{n\to\infty} \frac{1}{n}\log(M_n^2(\pi_n)) \quad &\geq \quad -[x_\xi\cdot\xi - \Lambda_n(x_\xi)] - \varepsilon\|x_\xi\| \\ &\qquad - \lim\sup_{n\to\infty} \frac{1}{n}\log[\sum_{i=1}^m \alpha_i [e^{n\ (\varepsilon\|x_{\beta_i}\| + x_{\beta_i}\cdot\xi - \Lambda_n(x_{\beta_i}))}]^{-1}] \\ &= \quad -\Lambda^*(\xi) - \varepsilon\|x_\xi\| - \max_{i=1,2,\ldots,m} \varepsilon\|x_{\beta_i}\| + x_{\beta_i}\cdot\xi - \Lambda_n(x_{\beta_i}) \end{aligned}$$

Letting $\varepsilon \to 0$, we get:

$$\lim\inf_{n\to\infty} \frac{1}{n}\log(M_n^2(\pi_n)) \quad \geq \quad -\Lambda^*(\xi) - \max_{i=1,2,\ldots,m} + x_{\beta_i}\cdot\xi - \Lambda_n(x_{\beta_i})$$

This is true for any $\xi \in B_v^o \cap D^o$. We assumed that $\bar{B}_v = \bar{B}_v^o$. Also we used $B_v \cap D^o \neq \emptyset$. And all the minimum rate points are in D^o. Thus we can obtain the minimum rate point as a limit of points $\xi \in B_v^o \cap D^o$. Since Λ^* is lower-semicontinuous. The above inequality folds if ξ is one of the minimum rate points. Since for any $b \in R^d$, $b\cdot\xi - \Lambda_n(b) \leq \Lambda_n^*(\xi)$ and equality holds if and only if $b = x_\xi$. Let us choose maximum β_j such that $x_{\beta_j}\cdot\xi - \Lambda_n(x_{\beta_j}) = \max_{i=1,2,\ldots,m}\{x_{\beta_j}\cdot\xi - \Lambda_n(x_{\beta_j})\}$. Then we have $-\Lambda^*(\xi) - (x_{\beta_j}\cdot\xi - \Lambda_n(x_{\beta_j})) \geq -2\Lambda^*(\xi) \geq -2\Lambda^*(B_v)$. The equality holds if and only of $\beta_j = \xi$ which is the MRP. Therefore, if $\xi \notin \beta$, we have the strict inequality, and therefore π_n can not be asymptotically efficient.

As Bucklew ([6]) points out, if we miss few local minima in creating π^*, the twist may sill be asymptotically efficient if the contributions from those missed local minima are small or not important. Furthermore, even if the set of local minima may not be finite or B_v is concave, $\forall\varepsilon > 0$, one can always find a finite numbers of local minima $\{\bar{\beta}_1, \bar{\beta}_2, , , , , \bar{\beta}_m\}$ such that $\bar{B}_v \subset \cup_{i=1}^m H_{\bar{\beta}_i}$, and $\Lambda^*(\bar{\beta}_i) \geq \Lambda^*(B_v) - \varepsilon$. Therefore, one still have the relation: $\lim\sup_{n\to\infty} \frac{1}{n}\log(M_n^2(\pi_n)) \leq -2(\Lambda^*(B_v) - \varepsilon)$. Thus, the asymptotical efficiency is obtained by letting $\varepsilon \to 0$. The importance twist can be constructed similarly as:

$$\pi^*(x) = f(x)\sum_{i=1}^m \alpha_i e^{x_{\bar{\beta}_i}\cdot x - \Lambda(\bar{\beta}_i)}, \ \ \sum_{i=1}^m \alpha_i = 1 \tag{3.20}$$

Chapter 4

Numerical Results (I.I.D. Case)

4.1 Results: Simple Portfolio

At first we tested simple portfolios for 2 dimensional cases (meaning 2 risk factors using regular stock and an in-the-money call option) and 8 dimensional cases (8 risk factors with various combinations of short/long positions of in-the-money/out-of-the-money call/put options and regular stocks). The detailed descriptions of these portfolios are found in Appendix D, and tables for these test results are located in Appendix E. In all cases, the risk factors are assumed to be correlated with each other. Tests were performed for various confidence levels, and holding periods to see how each method (importance sampling versus control variate) contributed to the reduction of errors. We performed non-linear optimization with both Gaussian and exponential twists. For all these tests, initial mean candidates are computed using the GHS method (See Appendix B). Then, the non-linear optimization is used to locate all local minima. We also performed optimization with several initial guesses which are different from the GHS result. Our tests showed that the optimization is robust, locating local minima quickly without using the GHS method. We shall use variance ratio to study the effectiveness of each method. The variance ratio is given by : $\frac{\sigma^2_{<Reg.MC>}}{\sigma^2_{<Each.MC>}}$ where $\sigma^2_{<Reg.MC>}$ is the variance for the regular Monte Carlo method, and $\sigma^2_{<Each.MC>}$ is the variance for each Monte Carlo technique applied. The higher variance ratio means better performance. In this section, we shall define the term 'the extreme quantile' to mean $p \leq 0.01$, and 'the non-extreme quantile' to mean $p > 0.01$ (Typically $p = 0.05$). We shall also define the term 'the longer holding period' to mean more than 10 days, and 'the shorter holding period' to mean less than 10 days (Typically 1 day).

The importance sampling method using non-linear optimization generally performed well in reducing errors for all cases. It works much better for the extreme quantile cases as expected from the Large Deviation Theory. Variance were reduced by a factors ranging from 22 to 50 (Table 1: Case: 1, 3, 5, 7) for both the GHS twist and our Gaussian twist. Since these portfolios have only one local minimum and relatively linear, the closeness of the GHS twist and our Gaussian twist is expected. Especially for portfolio with very short horizon (Table 1: Case: 5, 6), our experiments showed almost identical variance ratios for both twists. The exponential twist exhibited variance ratios ranging from 120 to 200 for all extreme quantile cases (Table 1: Case: 1, 3, 5, 7). This is expected from our reasoning given in the previous sections. The control variate in combination with the importance sampling method contributed very little to the extreme quantile cases (Table 2: Case: 1, 3, 5, 7). This makes sense since the control variate does not work well for the rare event simulation.

For the non-extreme quantile cases, the importance sampling techniques becomes less effective (but still better than the regular Monte Carlo). The variance ratios for the Gaussian twist decrease by factors ranging from 2 to 10 (Table 1: Case: 2,4,6,8), and the variance ratios for the exponential twists drop to 2 digit range (13 to 40). This is to be expected from the theory as well. Yet, even for the non-extreme quantile cases, the exponential twist maintains variance ratio of over 2 digits. On the other hand, the control variate method when taken together with the importance sampling techniques boosted variance ratios for both the Gaussian twist and the exponential twist by almost 2 to 8 times (Table 2: Case:2,4,6,8). This is so because the control variate method becomes more effective as the quantile becomes non-extreme. Therefore, for the extreme quantile cases, the importance sampling method alone performs the best. For the non-extreme quantile cases, the combination of the importance sampling and the control variate method seems to perform best.

4.2 Results: Complex Portfolio

In this section, we shall describe experimental results for more complex portfolios. Similar to the previous section, we shall demonstrate the effectiveness of our new exponential twist density with non-linear optimization. The exponential twist will be compared against both our Gaussian twist and the Gaussian twist by the GHS method (Appendix B). The precise descriptions of these portfolios can be found in Appendix E (Tables 2, 3). In these tests, we computed 10-day 99% VAR probability ($p \sim 0.01$) from 10 dimensions to 100 dimensions both correlated and uncorrelated.

Our Gaussian twist seems to perform similarly with the GHS method for many portfolios with one local minimum. However, for delta-hedged portfolios with multiple local minima, our Gaussian twist consistently provided variance ratio larger than 12 where the GHS method provided variance ratio below 10 (Table 2). The GHS method performed less (variance ratio $= 0.34 < 1$) than the regular Monte Carlo method for a portfolio with DAO (Down-And-Out) calls delta hedged with CON (Cash-Or-Nothing) puts. This portfolio is discontinuous and its delta is zero, so the Taylor approximation used by the GHS method becomes inaccurate. This point was mentioned by the GHS paper as well ([41]). On the other hand, our Gaussian twist still gave 2 digit variance ratio (1.213E+01). This experiment demonstrated that our method succeeded in locating multiple local minima, and creating twist based on these points. We performed tests on 100 dimensional portfolio as well and obtained similar results (Table 2).

Our experiments confirmed that our exponential twist outperformed both our Gaussian twist and the GHS method by factors ranging from 2 to 30 for all portfolios including ones with multiple local minima. Variance ratios for the exponential twist are anywhere from 22 to 400. In some cases, it went over tripple digits (1.12E+03). It appears that the exponential twist resulted in near tripple digit variance ratios for many portfolios. Our experiments support our claim in the previous chapter that the exponential twist should outperform Gaussian counterpart at least by a factor of 2. Therefore, the exponential twist combined with non-linear optimization is an excellent method for rare event simulation.

Chapter 5

New Importance Sampling Technique using Markov Chain Monte Carlo

5.1 Introduction

In many applications, it is not feasible to compute the maximum of the probability density $f(x)$, over the VAR region $V \leq v$. Particularly for the path dependent options, the dimension of x may be high and the formula for f may be complicated. We explore a different importance sampling strategy for such cases that only requires us to evaluate $V(x)$ and sample from the density $f(x)$ one component at a time (See below). The twisted density we sample is:

$$\pi_\theta(x) = C(\theta) \cdot f(x) \cdot e^{-\theta \cdot \max(V(x), v)} \tag{5.1}$$

where f is a correlated standard normal distribution, V is a portfolio value, v is a given VAR, and C is a normalization constant. The exponential twist $f(x) \cdot e^{-\theta V(x)}$ is common in many specific large deviation results, but there seems to be no theoretical justification for using it in the generality we do. Replacing $e^{-\theta V(x)}$ with $e^{-\theta \cdot \max(V(x), v)}$ avoids potentially unbounded $\pi_\theta(x)$. For example, take $S = S_0 e^{\sigma\sqrt{t}x - \frac{1}{2}\sigma^2 t}$ where $x \sim N(0,1)$, and $V(S) = -S$. Then the density $ce^{-\frac{1}{2}x^2} e^{\theta S_0 e^{\sigma\sqrt{t}x - \frac{1}{2}\sigma^2 t}}$ would be unbounded as $x \to \infty$. The twist formula we use pulls x toward the VAR region only until the VAR region is reached. In practice, the potential unboundedness of π_θ is a real problem. Furthermore, this twist has the desired effect near $V = v$. Take for example a simple case $V = -x$ and $v = -x_0$. Then $\pi_\theta(x) = C \cdot e^{-\frac{x^2}{2} + \theta x} = C \cdot e^{-\frac{(x-\theta)^2}{2} + \frac{\theta^2}{2}}$, for $x \leq x_0$, and $\pi_\theta(x) = C \cdot e^{-\frac{x^2}{2} + \theta x_0} = C \cdot e^{-\frac{(x-\theta)^2}{2} - \theta x + \theta^2 - \frac{\theta^2}{2} + \theta x_0}$ for $x > x_0$. If we let $\theta = x_0$, then we have $\pi_\theta(x) = C \cdot e^{-\frac{(x-x_0)^2}{2} + \frac{x_0^2}{2}}$ for $x \leq x_0$, and

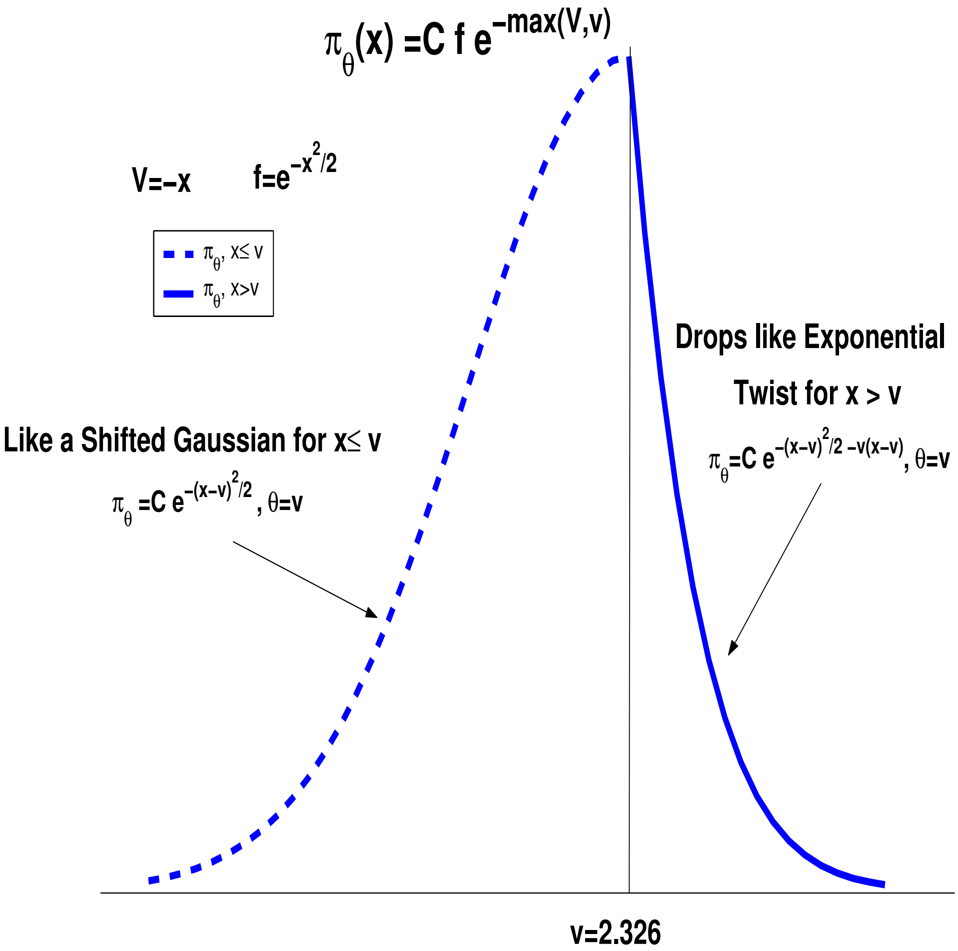

Figure 5.1: Density Plot: π_θ for $\theta = v = 2.326$ where $f = e^{-\frac{1}{2}x^2}$ and $V = -x$. π_θ looks like a shifted Gaussian twist for $V \leq v$, and drops like a shifted exponential twist for $V > v$. Adaptive MCMC will pick optimal θ so that $E_\theta[sign(V - v)] = 0$. The optimization finds $\theta = 4.137$.

$\pi_\theta(x) = C \cdot e^{-\frac{(x-x_0)^2}{2} - x_0(x-x_0) + \frac{x_0^2}{2}}$ for $x > x_0$. The striking fact about this twist is that it is behaving like a shifted Gaussian twist for $x \leq x_0$, and drops like a shifted exponential twist for $x > x_0$ (See Fig (5.1)). Thus, we can see the analogos connection of this twist with the twists we explained in the previous chapters.

This density $\pi_\theta(x)$ is parameterized with θ which is unknown, and can not be sampled by independent sampling. Therefore, We shall use Markov Chain Monte Carlo (MCMC) to sample from this new twist π_θ. We also propose an algorithm which combines the Metropolis-Hastings $M(R^2T)$ algorithm and the Robbins-Monro algorithm to sample from π_θ with an optimal θ so that the $E_\theta[sign(V - v)] = 0$. The idea is to to find an optimal target density π_θ so that most samplings are done near the VAR region where $V = v$. The Robbins-Monro algorithm is a stochastic approximation method. Combining MCMC with stochastic approximation is often called an Adaptive Markov Chain Monte Carlo. In the next section, we shall explain the Metropolis-Hastings $M(R^2T)$ algorithm.

5.2 Metropolis Hasting $M(RT^2)$ Algorithm

The Metropolis Hasting $M(RT^2)$ algorithm is a rejection technique for sampling from a probability density π which can not be sampled independently. This sampling is done under the Markov Chain Monte Carlo (MCMC) framework. In setup of MCMC algorithm, Markov chains are constructed from a transition kernel $T(x|y)$ which is a transition probability density satisfying: $P(X_{n+1} \in A|X_n = x) = \int_A T(x|y)dy, \quad A \subset R^d$. The sampling of X_{n+1} depends only on X_n, and does not depend on any earlier $X_{i\ i<n}$. That is $Pr(X_{n+1} \in A|X_n = x) = Pr(X_{n+1} \in A|X_n = x, X_{n-1} = x_{n-1}, ..., X_0 = x_0)$. MCMC method is any method producing an ergodic Markov chain (X_{n+1}) whose stationary distribution is π. The word 'ergodic' means that the law of large numbers hold:, i.e, $\frac{1}{N}\sum_{i=1}^{N} g(X_i) \approx \int_A g(x)\pi(x)dx$ where the limiting distribution of X_n is π. The Metropolis Hasting $M(RT^2)$ algorithm guarantees the ergodicity and its limiting distribution as π under some specific condition on T. These are irreducibility (T allows free moves all over the state A), and aperiodicity (The chain does not move cyclically). These notions will be defined next.

Definition 5.2.1 (Irreducibility)
A MCMC chain is said to be 'irreducible' if for any given $x \in A$, there exist $n \geq 1$ such that:

$$P(X_n \in A|X_0 = x) > 0 \tag{5.2}$$

This basically means that for each state in a set A, there is a positive probability to go to any other states in the set A and vice versa. If T is a strictly positive density, then the irreducibility is automatically satisfied.

Definition 5.2.2 (Aperiodicity)
Let us define the 'period of $x \in A$' to be the largest common divisor of the numbers $n > 0$ such that

$$P_x^n \equiv P(X_n = x|X_0 = x) > 0 \tag{5.3}$$

If a MCMC chain is irreducuble and $n = 1$, then the chain is said to be 'aperiodic'.

The above P_x^n is often called the 'n-step transition probability' (Sokal [64]). This basically says that the move from one state to another state is always possible, but never periodically. By Roberts and Tweedie ([62]), if the target density π is bounded and positive on every compact set of its support A, and if there exist positive numbers ε and δ such that $q(y|x) > \varepsilon$ if $|x - y| < \delta$. Then the Metropolis Hasting $M(RT^2)$ chain is irreducible and aperiodic. Typically one can take q to be a Gaussian distribution and the above condition is satisfied.

The $M(RT^2)$ algorithm satisfies the detail balance condition which implies the stationarity. We will define the detail balance condition, and show that the $M(RT^2)$ algorithm satisfies it.

Definition 5.2.3 (Detail Balance)
A MCMC chain with a transition probability T is said to satisfy the detail balance condition for the target density π if the following condition is satisfied:

$$T(x|y) \cdot \pi(x) = T(y|x) \cdot \pi(y), \quad \forall x, y, \in R^d, x \neq y \tag{5.4}$$

For the $M(RT^2)$ algorithm, $T(x|y)$ is given as:

$$\begin{aligned} T(x|y) &= a(x|y)q(y|x) + \delta_x(y)(1 - \int_A a(x|y)q(y|x)dy) \\ &= a(x|y)q(y|x) + \delta_x(y) \int_A (1 - a(x|y))q(y|x)dy \end{aligned} \tag{5.5}$$

where $q(y|x)$ is a proposal density of y given x, $\delta_x(y)$ is a Dirac mass in x and a is an acceptance probability given by:

$$a(x|y) = 1 \wedge \frac{q(x|y)\pi(y)}{q(y|x)\pi(x)} \tag{5.6}$$

If $\frac{q(x|y)\pi(y)}{q(y|x)\pi(x)} > 1$, then $a(x|y)q(y|x)\pi(x) = 1 \cdot q(y|x)\pi(x) = \frac{q(y|x)\pi(x)}{q(x|y)\pi(y)} q(x|y)\pi(y) = a(y|x)q(x|y)\pi(y)$. If $\frac{q(x|y)\pi(y)}{q(y|x)\pi(x)} \leq 1$, the similar calculation shows $a(x|y)q(y|x)\pi(x) =$

$a(y|x)q(x|y)\pi(y)$. Similarly, we have $\delta_x(y)(1-\int_A a(x|y)q(y|x)dy)\pi(x) = \delta_y(x)(1-\int_A a(y|x)q(x|y)dx)\pi(y)$. Therefore, for the $M(RT^2)$ algorithm, the detail balance condition is satisfied. Let us define the notion of stationarity.

Definition 5.2.4 (Stationarity)
A MCMC chain with a transition probability T is said to satisfy the stationarity condition if the following condition is satisfied:

$$\int_A \pi(y)T(y|x)dy = \pi(x) \quad \forall x, y \in A \tag{5.7}$$

When detail balance is satisfied, then the stationarity condition is also satisfied since: $\int_A \pi(y)T(y|x)dy = \int_A \pi(x)T(x|y)dy = \pi(x)\int_A T(x|y)dy = \pi(x)$.
Therefore, with appropriate target density π (bounded from above and positive) with a Gaussian proposal density q, the Metropolis-Hastings $M(RT^2)$ satisfies the stationarity, the aperiodicity, and the irreducibility. Any MCMC chain which satisfies the above three conditions is ergodic. Thus both the law of large numbers and the CLT can be applied to compute the estimator and its error by simulation. The algorithm is described as follows.

Algorithm 5.2.5 (General Metropolis-Hastings $M(RT^2)$)

(1) Given X_n, sample Y_n from some proposal distribution $q(y|x = X_n)$

(2) Sample ξ from the uniform distribution on $[0, 1]$.

(3) Accept Y_n with probability $= \frac{q(X_n|Y_n)\pi(Y_n)}{q(Y_n|X_n)\pi(X_n)}$. That is, if $\xi < \frac{q(X_n|Y_n)\pi(Y_n)}{q(Y_n|X_n)\pi(X_n)}$, then accept the it and let $X_{n+1} = Y_n$.

(4) If step (3) is not satisfied, then reject it and let $X_{n+1} = X_n$.

(5) Repeat the above steps (1)-(4) for the rest of sampling.

In practice, like computing VAR probability, applying the $M(RT^2)$ algorithm for all coordinates together is very impractical. In the next sections, we shall develop another sampling methods which allow MCMC sampling one coordinate at a time.

5.3 Gibbs Sampling

The Gibbs sampling is a general MCMC technique for sampling from multidimensional densities proposed by Geman and Geman ([29]). The basic idea is to sample coordinatewide using conditional densities. For example, suppose we need to sample $\{X\} \in A \subset R^d$ with a density f with univariate conditional densities $f_1, f_2, ..., f_d$. This technique calls for sampling each X_i from a corresponding univariate density f_i while all other coordinates are fixed. That is to sample $X_i|x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_d \sim f_i(x_i|x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_d)$. This is also called the 'Partial Resampling' by Goodman and Sokal([43]). We shall describe sampling techniques which applies the Metropolis-Hastings $M(RT^2)$ componentwide for each coordinate within the Gibbs sampling framework in the coming sections.

5.3.1 Metropolis-within-Gibbs Algorithm

The Metropolis-within-Gibbs Algorithm is a algorithm where the one dimensional Metropolis algorithm is applied one coordinate at a time. This is term is first defined by Tierney([66]). This may sound odd to some readers since the Gibbs sampling is a special case of the Metropolis algorithm. The distinction between the general Gibbs sampling and the Metropolis-within-Gibbs algorithm is that the former samples each X_i independent of the old X_i with acceptance probability =1. Whereas, the latter samples each X_i given the old value of X_i with some acceptance probability a. There are two versions of Metropolis-within-Gibbs algorithm. One method is called the Sequential Scan Metropolis-within-Gibbs, This is to sample each coordinate in fixed order for each iteration. The other method is called the Random Scan Metropolis-within-Gibbs. In this method, coordiantes are sampled randomly for each dimension. The next sections will describe these methods in more detail.

5.3.2 Sequential Scan Metropolis-within-Gibbs Algorithm

The Sequential Scan Metropolis-within-Gibbs is a Metropolis-within-Gibbs type sampling algorithm. Assuming the strict positivity of the target density π, and the proposal density q as in the previous section (this guarantees the aperiodicity and the irreducibility). Let us assume that the state is d dimensional with some given initial values $X = \{X_1^1, X_1^2, ...X_1^d\}$. Starting with the first coordinate X_1^1, we sample the proposal candidate Y_1^1 given X_1^1 while all other coordinate is being fixed with its proposal density given by $q(y|x = X_1^1)$. Next, we implement one dimensional Metropolis type algorithm to update the first coordinate by X_2^1 with while all other coordinates are fixed. Then we sample the second coordinate Y_1^2 given X_1^2 with proposal density $q(y|x = X_1^2)$. Similar to the first coordinate update, we update the second coordinate X_2^2 by with Metropolis type algorithm. We continue like this until the update of the last coordinate X_2^d for the first run. Thus, we repeat the above one-coordinate-at-time update for the rest of the simulation. Schematically, we have the following updating schedule:

$$\begin{aligned}
(X_1^1, X_1^2, ..., X_1^d) \\
(X_2^1, X_1^2, ..., X_1^d) \\
(X_2^1, X_2^2, ..., X_1^d) \\
...... \\
(X_2^1, X_2^2, ..., X_2^d) \\
(X_3^1, X_2^2, ..., X_2^d) \\
...... \\
(X_n^1, X_n^2, ..., X_{n-1}^d) \\
(X_n^1, X_n^2, ..., X_n^d)
\end{aligned} \tag{5.8}$$

Formally, the algorithm is described as follows:

Algorithm 5.3.1 (Sequential Scan Metropolis-within-Gibbs Algorithm)

(1) Start with the first coordinate $k = 1$.

(2) Given X_n^k*, sample* Y_n^k *from some proposal distribution* $q(y|x = X_n^k)$ *while all other coordinates are being fixed.*

(3) Sample ξ *from the uniform distribution on* $[0, 1]$.

(4) Accept Y_n^k *with probability* $= \frac{q(X_n^k|Y_n^k)\pi(Y_n^k)}{q(Y_n^k|X_n^k)\pi(X_n^k)}$*. That is, if* $\xi < \frac{q(X_n^k|Y_n^k)\pi(Y_n^k)}{q(Y_n^k|X_n^k)\pi(X_n^k)}$*, then accept the it and let* $X_{n+1} = (X_n^1, ...Y_n^k, ...X_n^d)$.

(5) If step (4) is not satisfied, then reject it and let $X_{n+1} = (X_n^1, ...X_n^k, ...X_n^d)$.

(6) Increment the next k by 1 sequentially ($k = k + 1$)*, and the repeat the above steps (1)-(5) until all d coordinates are sampled.*

(7) Repeat the above steps (1)-(6) for the entire loops.

Notice that the significant part of this algorithm is the steps (4) and (5) above. Whether one coordinate is accepted or rejected, other fixed coordinates are reported together. This means that each acceptance/rejection of just one coordinate of a vector results in accepting or rejecting the entire vector.

With the right target density and proposal density, the Sequential Metropolis-within-Gibbs method satisfies the aperiodicity and the irreducibility. However, the detail balance condition is not satisfied. The d-dimensional Sequential Metropolis-within-Gibbs Transition operator is a product operator $T = \prod_{i=1}^d T_i$. Each T_i acting on the i-th coordinate satisfies the detail balance condition, but the product T does not. Nevertheless, the stationarity is satisfied. Therefore, the ergodicity holds, and this algorithm is guaranteed to work. There is another way of performing the coordinatewide Metropolis-within-Gibbs. This is called the Random Scan Metropolis-within-Gibbs.

5.3.3 Random Scan Metropolis-within-Gibbs Algorithm

The Random Scan Metropolis-within-Gibbs algorithm is another type of the Metropolis-within-Gibbs sampling method. This algorithm still performs the componentwide Metropolis-within-Gibbs algorithm one coordinate at a time, but randomly. So the algorithm is given as follows:

Algorithm 5.3.2 (Random Scan Metropolis-within-Gibbs Algorithm)

(1) Given numbers $\bar{k} = 1...d$, randomly shuffle $\bar{k}$ to get new set of coordinate array $\{\hat{k}\} = randperm(\{\bar{k}\})$.

(2) Start with the first index $\bar{k} = 1$, pick the randomly chosen coordinate by $k = randperm(\bar{k})$.

(3) Given X_n^k, sample Y_n^k from some proposal distribution $q(y|x = X_n^k)$ while all other coordinates are being fixed.

(4) Sample ξ from the uniform distribution on $[0, 1]$.

(5) Accept Y_n^k with probability $= \frac{q(X_n^k|Y_n^k)\pi(Y_n^k)}{q(Y_n^k|X_n^k)\pi(X_n^k)}$. That is, if $\xi < \frac{q(X_n^k|Y_n^k)\pi(Y_n^k)}{q(Y_n^k|X_n^k)\pi(X_n^k)}$, then accept the it and let $X_{n+1} = (X_n^1, ...Y_n^k, ...X_n^d)$.

(6) If step (5) is not satisfied, then reject it and let $X_{n+1} = (X_n^1, ...X_n^k, ...X_n^d)$.

(7) Increment the next k by 1 sequentially ($k = k + 1$), and the repeat the above steps (1)-(6) until all d coordinates are sampled.

(8) Repeat the above steps (1)-(7) for the entire loops.

The algorithm is the same as the Sequential Metropolis-within-Gibbs except for the order of updates. Experimentally, the performance is quite similar to the sequential counterpart. However, this has the theoretical advantage in that there are many papers (Rosenthal [61]) showing that this chain is not only ergodic, but also geometrically ergodic for many target densities (including our π_θ, section 6.2.2) quite easily. While the transition operator for the Sequential Scan Metropolis-within-Gibbs is a product operator, the transition operator for the Random Scan Metropolis-within-Gibbs is a summation operator $T = \frac{1}{d}\sum_{i=1}^{d} T_i$. It is quite easy to see that the total transition operator satisfies the detail balance condition since each T_i does. Therefore, the stationarity is automatically satisfied.

5.3.4 Symmetric Random-Walk Metropolis

For the Metropolis algorithms, the choice of proposal density is a very important issue. Beside the fact that the target density π has to be absolutely continuous with respect to the proposal density q, one must carefully chose q so that it is effective. One kind of proposal density that guarantees the aperiodicity and the irreducibility very well is the symmetric random-walk density. It is in generally any strictly positive function defined in X satisfying $q(x|y) = q(y|x)$. The typical candidate is given by $q(x|y) = q(|x-y|)$ where q is either the correlated Gaussian density or the uniform density. It is clear from the formulation that it is a good symmetric random-walk density. Use of the random-walk density is appealing both from the implementation point of view and the theoretical point of view. From the implementation point of view, the expression for the acceptance probability is simplified as:

$$\frac{q(X_n|Y_n)\pi(Y_n)}{q(Y_n|X_n)\pi(X_n)} = \frac{\pi(Y_n)}{\pi(X_n)} \tag{5.9}$$

Therefore this fact should contribute computationally. We shall use this symmetric random-walk proposal density for our new algorithm in the next section.

From the theoretical point of view the symmetric random-walk algorithm simplifies the expression for the Metropolis transition probability kernel. Remember that the Metropolis kernel is written in general as:

$$\begin{aligned} T(x|\cdot) &= \int_B q(y|x)a(x|y)dy \\ &\quad + I_B(x)\int_B q(x|y)(1-a(x|y))dy \end{aligned} \tag{5.10}$$

where $B \subset R$. The first expression is the probability of moving from x to any other states $y \in B$. The second expression is the probability of staying in the state x. Since q is symmetric, one can define the acceptance region $A(x) \subset B$ and the rejection region $R(x) \subset B$ easily as follows:

$$A(x) = \{y \in B : a(x|y) = \frac{q(x|y)\pi(y)}{q(y|x)\pi(x)} = \frac{\pi(y)}{\pi(x)} \geq 1\}.$$
$$R(x) = \{y \in B : a(x|y) = \frac{q(x|y)\pi(y)}{q(y|x)\pi(x)} = \frac{\pi(y)}{\pi(x)} < 1\}.$$

Then we can re-write the above expression (5.10) as:

$$\begin{aligned} T(x|\cdot) &= \int_{A(x)} q(|x-y|)dy + \int_{R(x)} q(|x-y|)a(x|y)dy \\ &\quad + I_B(x)\int_{R(x)} q(|x-y|)(1-a(x|y))dy \end{aligned} \tag{5.11}$$

These expressions will be used frequently in the next chapter. In the next section, we shall describe our new algorithm based on the Metropolis-within-Gibbs algorithm for the VAR probability computation.

5.4 New Algorithm for VAR Probability Computation

In this section we propose a new method that will let MCMC determine the best importance sampling twist through simulation. Our technique is motivated by the attempt to sample from a Gibbs distribution in statistical physics of the form $\pi = \frac{1}{Z}e(-E/T)$ where T is a temperature, E is an energy of a kinetic system, and Z is an unknown constant. This sampling has typically been done either by the Gibbs sampling or the Metropolis-within-Gibbs Algorithm. The implementation is very simple, and can be applied to any complex cases where non-linear optimization is not possible.

Our first invention is the introduction of the family of target densities (π_θ) parametrized by $\theta \in R^+$ as:

$$\pi_\theta = C \cdot f \cdot e^{-\theta \cdot \max(V,v)} \tag{5.12}$$

where C is a normalization constant, and f is a d-dimensional correlated normal density given by:

$$f = \frac{1}{(2 \cdot \pi)^{d/2} |\Sigma_x|} e^{-x^T \cdot \Sigma_x^{-1} \cdot x/2} \tag{5.13}$$

where Σ_x is a correlation matrix for the risk factors x. One can also translate the expression (5.12) by v to get the following equivalent format:

$$\begin{aligned} \pi_\theta &= C \cdot f \cdot e^{-\theta \cdot ((\max(V,v)-v)+v)} \\ &= \hat{C} \cdot f \cdot e^{-\theta \cdot \max(V-v,0)} \end{aligned} \tag{5.14}$$

One can see the striking similarities between the Gibbs density π and our target density π_θ. If $V - v \leq 0$, the density π_θ becomes larger. When $V - v > 0$, the density π_θ becomes smaller. So this density π_θ has the characteristics we are looking for. This will sample more in the region(s) of interest $V - v \leq 0$, and sample less in the other region(s). Thus sampling from π_θ is better than sampling from f if $\theta > 0$. However, θ can not be too big. Therefore some kind of optimization is called for. Another good feature of this formulation is that π_θ is bounded from above for any fixed θ. This is a very important point both from the theoretical view point (guarantees convergence) as well as from the practical view point. We found that the best θ^* satisfies the following condition:

$$E_{\theta^*}(sign(V - v)) = 0 \tag{5.15}$$

This mean that the expected numbers of incidents for $V - v \leq 0$ and the expected numbers of incidents for $V - v \geq 0$ are the same. This will results in the density almost peaking near the area where $V = v$. From the importance sampling point of view, this is what the non-linear optimization looks for. The idea is great. But how can we find such θ^* ? The answer lies in the Robbins-Monro algorithm from the stochastic approximation theory described in the next section.

5.4.1 Robbins-Monro (RM) Algorithm for VAR

We have proposed a good twist π_θ, but optimal θ need to be found so that (5.15) is satisfied. Another method we propose is to perform the following Robbins-Monro algorithm from stochastic approximation (SA) theory.

Algorithm 5.4.1 (Robbins-Monro Algorithm for VAR)
Let us assume that the portfolio function V is a continuous function defined in R^d. The Robbins-Monro Algorithm for VAR is a stochastic approximation algorithm for finding optimal θ^* satisfying (5.15) whose iterative process is given by:

$$\theta_{n+1} = \theta_n + \frac{1}{n} \cdot H(\theta, X_n) \tag{5.16}$$

where $H(\theta, x) = sign(V(x) - v)$, and X_n is sampled from the density π_θ (5.12) using the Metropolis algorithm.

The expression (5.16) is called the Robbins-Monro algorithm in general. There are many papers and text books on the Robbins-Monro algorithms in general ([52], [10], [11]). Typically, this algorithm is used to find such θ^* where the sampling is done independently. If the sampling is done independently, then the above algorithm finds the optimal θ^* so that (5.15) holds under some conditions (chapter 6). Recently, Andrieu ([1], 2002) proposed the combined approach of the Robbins-Monro algorithm and the Metropolis algorithm. His paper uses the Metropolis algorithm to sample from a fixed target density π while the proposal density q is optimized using the Robbins-Monro algorithm. However, for the VAR probability computation, the target density π_θ can not be independently sampled. The above algorithm calls for finding the optimal θ^* by the Robbins-Monro algorithm while sampling is done each time using the Metropolis algorithm. Application to the VAR probability computation in this manner is never undertaken before. We have found that the combined approach is possible, and verified both theoretically (chapter 6) and experimentally (chapter 8). The important part of this combined approach is that this results in sampling most of the time near the region where $V = v$, and appropriate optimal θ is found at the same time via simulation. Furthermore, this algorithm does not require computing gradient, or laplacian like other deterministic methods as in chapter 3. Therefore, it can be applied to any portfolios. We shall describe the actual algorithm in the next section.

5.4.2 The Combined Algorithm for VAR Probability Computation

So far, we have discussed the ideas behind the combined approach of the Robbins-Monro algorithm and the Metropolis algorithm. We shall describe the actual process in more detail. Staring with some small positive θ_0 (say 10^{-6}, or even 0), our approach is to sample from π_θ using the Metropolis-within-Gibbs Algorithm (either Random Scan or Sequential Scan) from the previous sections. Then update the next θ_{n+1} by $\theta_{n+1} = \theta_n + \frac{1}{n} \cdot sign(V(X_n) - v)$. This process is continued adaptively for the rest of the simulation. The Robbins Monro algorithm for VAR probability computation is given as follows:

Algorithm 5.4.2 (Effective VAR Probability Computation Algorithm)

(1) Start with the first step $n = 1$, Assign initial θ_n to be very small positive number (10^{-6} or even 0), and initial Gaussian risk factors $X_n = (X_n^1, X_n^2, ...X_n^d)$.

(2) Start with the first index or coordinate $k = 1$.

(3) Given X_n^k, sample Y_n^k from $q(y|x = X_n^k)$ while all other coordinates are being fixed. Here we use Gaussian symmetric random-walk proposal distribution with variance = 2.38 (To be explainend later).

(4) Sample ξ from the uniform distribution on $[0, 1]$.

(5) Accept with probability $\frac{q(X_n^k|Y_n^k)\pi(Y_n^k)}{q(Y_n^k|X_n^k)\pi(X_n^k)}$. That is if $\xi < \frac{q(X_n^k|Y_n^k)\pi(Y_n^k)}{q(Y_n^k|X_n^k)\pi(X_n^k)}$, then let $X_{n+1} = (X_n^1, ...Y_n^k, ...X_n^d)$.

(6) If step (5) is not satisfied, then reject it and let $X_{n+1} = (X_n^1, ...X_n^k, ...X_n^d)$.

(7) Update θ_{n+1} by $\theta_{n+1} = \theta_n + \frac{1}{n} \cdot sign(V - v)$. Make sure to record all the values of $\{\theta_n\}$, and $\{V_n\}$.

(8) Increment n by 1 ($n = n + 1$), and increment the next k by 1 sequentially or randomly ($k = k + 1$), and the repeat the above steps (2)-(7) until all d coordinates are sampled.

(9) Repeat the above steps (2)-(8) for the entire loops.

(10) When the entire loop is finished, then discard the first set of data (say 200) to eliminate initialization bias.

When the above sampling is finished, we have arrays of portfolio value $\{V_n\}$, and $\{\theta_n\}$. Suppose both have N points. The probability p can be computed by $p = \frac{\sum_{i=1}^{N} I_{V_i \leq v} \cdot e^{\theta_i \cdot max(V_n,v)}}{\sum_{i=1}^{N} e^{\theta_i \cdot max(V_n,v)}}$. Its error can be computed by the method of batched means as explained in the next section. If we want to find out VAR for a quantile $\hat{p}$, then just like as explained in chapter 2, sort $\{V_n\}$ in ascending order while keeping track of θ_n for each V_n. Then starting with $j = 1$, keep summing $\hat{p}^j = \frac{\sum_{i=1}^{j} e^{\theta_i \cdot max(V_i,v)}}{\sum_{i=1}^{N} e^{\theta_i \cdot max(V_n,v)}}$ until $\hat{p}^k = \hat{p}$. The $\hat{p}$-th quantile VAR is simple given by V_k. Note that we have an array of $\{\theta_n\}$, but they are nearly the same since θ_n converges to a unique limit θ^* (This will be proved in the next chapter).

So far, we have explained the algorithm; however we need to compute the simulation error based on a concept similar to the notion of standard deviation. For the i.i.d. cases, it is quite simpler (just $p \cdot (1-p)$). Therefore, the error is given by $\frac{\sqrt{p \cdot (1-p)}}{\sqrt{N}}$. For the MCMC cases, one can perform the similar tasks using the Method of Batched Means. This will be explained in the last section of this chapter.

5.4.3 Trick for Effective Dimension Reduction and Implementation

The above Metropolis-within-Gibbs algorithm sweeps through each coordinate one at a time. However, the target density π_θ involves computing the portfolio values V for each Metropolis step. Typically V is a linear combinations of individual securities or derivatives $V_i(x_i)$ each of which depends on one underlying risk factors x_i (unless path dependent options are involved). Therefore $V = \sum_{i=1}^{d} C_i V_i(x_i)$ where C_i is a constant which represents the numbers of positions plus (long) or minus (short) for each security V_i. For the sake of simplicity, let us assume that portfolio does not contain any path-dependent options (this issue will be taken up later). Then we have the following target density formula π_θ:

$$\begin{aligned} \pi_\theta &= C \cdot f \cdot e^{-\theta \cdot \max(V,v)} \\ &= C \cdot f \cdot e^{-\theta \cdot \max(\sum_{i=1}^{d} C_i V_i(x_i), v)} \end{aligned} \tag{5.17}$$

where f is given by:

$$f = \frac{1}{(2 \cdot \pi)^{d/2} |\Sigma_x|} e^{-x^T \cdot \Sigma_x^{-1} \cdot x/2} \tag{5.18}$$

with correlation matrix Σ_x. Then the acceptance probability $a_{\theta_n}(X_n^k|Y_n^k) = \frac{q(X_n^k|Y_n^k)\pi_{\theta_n}(Y_n^k)}{q(Y_n^k|X_n^k)\pi_{\theta_n}(X_n^k)}$ can be computed as follows:

$$
\begin{aligned}
a_{\theta_n}(X_n^k|Y_n^k) &= \frac{\pi_{\theta_n}(Y_n^k)}{\pi_{\theta_n}(X_n^k)} \\
&= \frac{f(Y_n^k) \cdot e^{-\theta_n \cdot \max(\sum_{i\neq k}^d C_i V_i(X_n^i) + C_k V_k(Y_n^k), v)}}{f(Y_n^k) \cdot e^{-\theta_n \cdot \max(\sum_{i\neq k}^d C_i V_i(X_n^i) + C_k V_k(X_n^k), v)}} \\
&= \frac{f(Y_n^k) \cdot e^{-\theta_n \cdot \max(D(t) + C_k V_k(Y_n^k), v)}}{f(Y_n^k) \cdot e^{-\theta_n \cdot \max(D(t) + C_k V_k(X_n^k), v)}} \\
&= \frac{e^{-\theta_n \cdot \max(D(t) + C_k V_k(Y_n^k), v)}}{e^{-\theta_n \cdot \max(D(t) + C_k V_k(X_n^k), v)}} \\
&\quad \cdot e^{-(x^T \cdot \Sigma_x^{-1} \cdot x/2)|_{x_k = Y_n^k} + (x^T \cdot \Sigma_x^{-1} \cdot x/2)|_{x_k = X_n^k}} \\
&= \frac{e^{-\theta_n \cdot \max(D(t) + C_k V_k(Y_n^k), v)}}{e^{-\theta_n \cdot \max(D(t) + C_k V_k(X_n^k), v)}} \\
&\quad \cdot e^{-\frac{1}{2} l_{k,k}((Y_n^k)^2 - (X_n^k)^2) - \sum_{i=1, i\neq k}^d l_{k,i} x_i (Y_n^k - X_n^k)}
\end{aligned} \tag{5.19}
$$

where $D(t) = \sum_{i\neq k}^d C_i V_i(X_n^i)$ and $l_{i,j}$ is the (i,j)-th element of the matrix Σ_x^{-1}. For our Metropolis-within-Gibbs algorithm, only one component is updated. Therefore only terms involving Y_n^k, and X_n^k need to be evaluated. Typically, computing the entire portfolio value V is very expensive. When implementing this algorithm, computing $D(t)$ is the expensive part of the algorithm since $D(t)$ involves $d-1$ of the total d dimensional securities V. Since $D(t)$ is the same as the previous step, we can just reuse the value, rather than recomputing it. This simple trick will reduce the computational time by a factor of d (dimension). Therefore, the larger the dimension, the more efficient this implementation will be.

The above method work for many typical portfolios (for both correlated and uncorrelated problems). However, we have encountered several problems for highly correlated portfolio with many local minima. For these special cases, we found that special trick is needed to counter this problem. The method we recommend for such unusual cases is to implement a kind of Principal Component Axis approach. That is to use a correlated symmetric-random walk proposal density whose correlation matrix is the same as the correlation matrix Σ_x of the target density π_θ. The idea is to take the Cholesky decomposition of the matrix Σ_x so that $C^T \cdot C = \Sigma_x$. When we generate a one dimensional candidate X_n^k, then we generate new set of vector by multiplying the new vector with the Cholesky factor C^T. That is to use $X = \{X_n^1, ...Y_n^k, X_{n-1}^{k+1}, ...X_{n-1}^d\}$ to get new $\bar{X}$ by $\bar{X} = C^T \cdot X$. This has the effect of diagonalizing the density f. The idea is great, and it works for almost all cases (except some even more unusual cases). The only catch is that the componentwide evaluation of V described above can not be done here. Since new $\bar{X}$ contains the one dimensional candidate Y_n^k for all dimension of the vector because of the Cholesky multiplication. This means that the effect of this proposal candidate is echoed to all component of V. Thus, one has to evaluate V for each componentwide update of underlying risk factors. Experiments have shown that even in this case, we have seen this method outperforming the plain Vanilla Monte Carlo by a factor of 2 to 4. One could also decompose V into several groups of securities where the correlation between each group is very small if not zero. When that is done, then one can still implement the similar blockwide update for each groups. If there are n such groups, then this grouping should decrease the computational cost by $\frac{d}{n}$. Furthermore, for the path-dependent options, one can treat each correlated risk factors into several independent increment by the nature of Brownian motion. Therefore, this kind of trick should also aid us in reducing calculation dimension for each update.

So far, we have discussed several sampling tricks for possible dimension reduction. Another trick we recommend is to use '2.38' as the variance for the symmetric-random walk proposal density. This point was pointed out by Gilks([30]), and our experiments have shown better convergence behavior than any other parameters we have used. One could also use the Robbins-Monro type algorithm to optimize this parameter as being done by Andrieu ([1]). In his paper, the proposal density is optimized for a fixed target density. In our case, the approach is opposite. The target density is optimized for a fixed random-walk proposal density. Conceptually one could try to optimize both the target and the proposal densities, however, we have not experimented this yet. This can be left as a future topic of research. Furthermore, we have applied Chen's algorithm (to be explained in the later chapter) which is easier to implement for our various test portfolios (chapter 8). This is implemented to control the behavior of θ during our Robbins-Monro process. This algorithm is explained in the next section so that our main algorithm in the previous section does not get overcrowded with many details.

5.5 Error Analysis

For the i.i.d. case, the error for an estimator $\hat{y} = \sum_{i=1}^{N} \frac{Y_i}{N}$ is given by expression of the form $\sqrt{\frac{\sum_{i=1}^{N} \frac{Y_i^2}{N} - (\sum_{i=1}^{N} \frac{Y_i}{N})^2}{N}}$ which is guaranteed by the CLT. For the MCMC cases, the similar statement can be made if the chain is ergodic. This is a MCMC version of the law of large numbers. As we will show in the next chapter, our Metropolis-within-Gibbs algorithm is ergodic (actually geometrically ergodic). Therefore, error analysis can be made similar to the i.i.d cases. For the MCMC cases, there is a notion of autocorrelation time. This is a measure of how strongly each generated samples are correlated. Smaller the autocorrelation time τ, the better or more independent the chain is. In the next section, the concept of autocorrelation time will be explained.

5.5.1 Autocorrelation Time

The autocorrelation time τ tells how many τ points are effectively correlated among N sample points. In i.i.d. case $\tau = 1$ by definition. For the MCMC cases, only $\frac{N}{\tau}$ points are effectively independent. Therefore smaller the τ, the more independent samples are available for the same N. So $\tau > 1$ for MCMC in general. There are many attempts to calculate this quantity both from theoretical point of view and implementation point of view. We shall discuss briefly about the concept of τ. The detail of this subject can be found in Sokal[23]. There are two formulations of autocorrelation time. One is called the exponential autocorrelation time, and the other is called the integrated autocorrelation time. The former is a more theoretical concept defined in the Hilbert space $l^2(\pi)$ which is a space of complex-valued square-integrable functions with respect to a density π with the norm $\|f\| = (\int_x |f(x)|^2\pi(x)dx)^{1/2} < \infty$. In general, the exponential autocorrelation time $\tau_{exp} \geq \tau_{int}$, and τ_{exp} provides the upper bound for analysis purpose. While τ_{int} provides more concrete information about MCMC chains. We can make few observation about our Metropolis chain using the concept of τ_{exp}. For the transition operator T for our Metropolis, '1' is the only (unique) eigenvalue on the unit circle (corresponding to the steady state distribution π which is the eigenvector) since our chain is aperiodic. Furthermore, if R is the spectral radius of T acting on the orthogonal complement of the spectra function, then the exponential autocorrelation time τ_{exp} is related to R by $R = e^{-1/\tau_{exp}}$. If λ_{max} is the sup of the spectral radius other than 1 in R, then $\tau_{exp} = \frac{-1}{\log \lambda_{max}}$. This is a positive number since all other eigenvalues are strictly less than 1 for our chain.

Let us define the integrated autocorrelation time τ_{int}. This is more practical since we can either compute explicitly (for some simple cases), or estimate via simulation for all cases after steady state is being reached. The integrated autocorrelation time $\tau_{int} = \tau$ is defined as:

$$\tau = \sum_{t=-\infty}^{\infty} \rho(t) = 1 + 2 \cdot \sum_{t=1}^{\infty} \rho(t) \tag{5.20}$$

where $\rho(t) = \lim_{k\to\infty} corr(Y_k, Y_{k+t}) = \lim_{k\to\infty} \dfrac{cov(Y_k, Y_{k+t})}{\sigma(Y_k) \cdot \sigma(Y_{k+t})} \subset [0, 1]$. In the next section, the integrated autocorrelation will be computed via simulation.

5.5.2 Method of Batched Means: Integrated Autocorrelation Time

For some simple model problems, the autocorrelation time can be computed analytically. However, for general MCMC problems, it can not be computed analytically like our VAR calculation. The method to estimate this quantity is called the Method of Batched Means. The idea is to break up the entire N samples into the batch size m, and the number of batches b so that $N = m \cdot b$. If the batch size is large enough, then the dependence will be contained within each batch. Suppose we want to estimate the error for computing the estimator $\hat{y} = \sum_{i=1}^{N} \frac{Y_i}{N}$. If $\{Y_n\}$ are created via MCMC simulation, then there is an autocorrelation between them. Let us describe how one can estimate an error for this chain. At first compute the total variance pretending that the MCMC is i.i.d. as:

$$Var(Y) = \frac{1}{b \cdot m} \sum_{i=1}^{N} Y_i^2 - (\frac{1}{b \cdot m} \sum_{i=1}^{N} Y_i)^2 \tag{5.21}$$

Then next, compute the variance of the each 'Batched Mean' as:

$$\begin{aligned} Var(Y^m) &= \frac{1}{b-1} \sum_{i=1}^{b} (\frac{\sum_{k=m \cdot (i-1)+1}^{m \cdot i} Y_k}{m} - \frac{1}{b} \sum_{i=1}^{b} \frac{\sum_{k=m \cdot (i-1)+1}^{m \cdot i} Y_k}{m})^2 \\ &= \frac{1}{b-1} \sum_{i=1}^{b} (\frac{\sum_{k=m \cdot (i-1)+1}^{m \cdot i} Y_k}{m})^2 - \frac{b}{b-1} (\sum_{i=1}^{b} \frac{\sum_{k=m \cdot (i-1)+1}^{m \cdot i} Y_k}{m})^2 \\ &= \frac{1}{b-1} \sum_{i=1}^{b} (\frac{\sum_{k=m \cdot (i-1)+1}^{m \cdot i} Y_k}{m})^2 - \frac{b}{b-1} (\sum_{i=1}^{b \cdot m} \frac{Y_i}{b \cdot m})^2 \end{aligned} \tag{5.22}$$

The effective batch size m_{eff} is given by $m_{eff} = \frac{Var(Y)}{Var(Y^m)}$. m_{eff} is also related to autocorrelation time τ via $m_{eff} = \frac{m}{\tau}$. Therefore, τ can be computed by:

$$\begin{aligned} \tau &= m \cdot \frac{Var(Y^m)}{Var(Y)} \\ &= m \cdot \frac{\frac{1}{b-1} \sum_{i=1}^{b} \left(\sum_{k=m \cdot (i-1)+1}^{m \cdot i} \frac{Y_k}{m} \right)^2 - \frac{b}{b-1} (\sum_{i=1}^{b \cdot m} \frac{Y_i}{b \cdot m})^2}{\frac{1}{b \cdot m} \sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} \frac{Y_i}{b \cdot m})^2} \\ &= m \cdot \frac{\frac{1}{b-1} \sum_{i=1}^{b} \left(\sum_{k=m \cdot (i-1)+1}^{m \cdot i} \frac{Y_k}{m} \right)^2 - \frac{b}{b-1} (\sum_{i=1}^{N} \frac{Y_i}{N})^2}{\sum_{i=1}^{N} \frac{Y_i^2}{N} - (\sum_{i=1}^{N} \frac{Y_i}{N})^2} \end{aligned} \tag{5.23}$$

Notice that this formula is easily implemented since only the expression $\sum_{i=1}^{b} \left(\sum_{k=m \cdot (i-1)+1}^{m \cdot i} \frac{Y_k}{m} \right)^2$ needs to be computed in addition to terms associated with computing error for the i.i.d. case. In our experiment, we set the batch size m to be anywhere from 500 to 1000. For the MCMC case, the effective variance can be computed as:

$$Var(Y)_{eff} = \frac{Var(Y)}{\tau} \tag{5.24}$$

The simulation error (95% confidence) can therefore be computed as:

$$Error(Y)_{eff} = 2\sqrt{\frac{Var(Y) \cdot \tau}{N}} \tag{5.25}$$

VAR probability involves ratio of two estimators $\frac{\frac{1}{N}\sum_{i=1}^{N} I_{V \leq v} L_i}{\frac{1}{N}\sum_{i=1}^{N} L_i} = \frac{\bar{Y}}{\bar{X}}$ where L_i is a likelihood ratio, and X and Y are the estimators for the numerator and the denominator respectively. For this case we can use the following formula for estimating the variance for $\frac{\bar{Y}}{\bar{X}}$:

$$Var(\frac{Y}{X})_{eff} = \frac{Var(Y)_{eff}}{\bar{Y}^2} + \frac{\bar{X}^2 Var(X)_{eff}}{\bar{Y}^4} - \frac{2\bar{X}Cov(X,Y)_{eff}}{\bar{Y}^3} \tag{5.26}$$

where $Var(Y)_{eff} = \frac{Var(Y)}{\tau_Y}$, $Var(X)_{eff} = \frac{Var(X)}{\tau_X}$, and $Cov(X,Y)_{eff} = \frac{Cov(X,Y)}{\tau_{X,Y}}$. τ_Y, τ_X, and $\tau_{X,Y}$ are the autocorrelation time for Y, X, and covariance of these respectively. The derivation of the above formula is based on Taylor expanding the ratio $\frac{Y+\Delta Y}{X+\Delta X}$. The calculation shall be skipped here as it is explained in Kendall and Stuart ([50]). The simulated error for $\frac{\bar{Y}}{\bar{X}}$ (95% confidence) is given as:

$$Error(\frac{Y}{X})_{eff} = 2\sqrt{\frac{Var(\frac{Y}{X})_{eff}}{N}} \tag{5.27}$$

5.5.3 Methods for Reducing Error due to Normalization Constant

The algorithm explained in the previous section may not show good convergence results for some cases because of relatively large variance for the normalization constant (or denominator for the ratio of two estimators). For this situation one can follow the following algorithm. This method is based on the Bayes rule (or conditional probability). Let us assume that we decide to run N iterations. We run $\frac{1}{5}$ of N iterations using the regular Monte Carlo, and the other $\frac{4}{5}$ of N iterations using the algorithm described in the previous sections (This is heuristic. One can use $\frac{1}{2}$ of N for each methods). Let us be more specific. At first, we run $\frac{N}{5}$ iterations of the regular Monte Carlo method, and compute VAR probability $p_1 = P(V < v_1)$ where $v_1 > v$ is chosen to approximately corresponds to about $p_1 = 0.1$ (or even $p_1 = 0.5$). This does not have to be chosen precisely. This Monte Carlo run gives the estimator $\hat{p}_1$ with the error $\varepsilon_1 = \sqrt{\frac{5(\hat{p}_1 - \hat{p}_1^2)}{N}} * 2$ (with 95 % confidence). Then we run the second $\frac{4N}{5}$ simulations using the adaptive MCMC algorithm from the previous section and compute the conditional probability $p_2 = P(V < v|V < v_1)$. Using the error estimate for the ratio of two estimators from the previous section, we get the probability estimate $\hat{p}_2$ with the effective error ε_2 as found by (5.27). The final VAR probability can be estimated by $\hat{p} = \hat{p}_1\hat{p}_2$. This is so because $\hat{p}_1\hat{p}_2 \approx P(V < v_1)P(V < v|V < v_1) = P(V < v_1)\frac{P(\{V<v\}\cap\{V<v_1\})}{P(V<v_1)}$ $= P(V < v_1)\frac{P(V<v)}{P(V<v_1)} = P(V < v)$ by the simple application of the Bayes rule. Thus the final error can be computed by $\varepsilon = \varepsilon_1\varepsilon_2 + \varepsilon_1\hat{p}_2 + \varepsilon_2\hat{p}_1$. The effectiveness of this approach is given by the fact that the first run is performed near the center of the distribution V so the estimate is reasonably accurate. The second conditional probability is more appropriate for the rare event scenario because it heavily samples near $V = v$. Furthermore, since it is a conditional probability given $V < v_1$, the events in $V > v_1$ are not sampled. The set $V > v_1$ typically cause a problem when estimating the normalization constant. Therefore, the second run reduces the variance for the normalization constant by throwing away many unwanted contributions to the variance for the denominator.

We have experimented with another formula for $\pi_\theta \sim fe^{\max(-\theta max(V-v,0),-C)}$ where $C \in [3,4]$ for smaller dimensions (≤ 10), and $C \in [2,3]$ for larger dimensions (> 15). This puts lower bound for the expression inside the exponent, thus effectively creates upper bound for the variation of the likelihood ratio. Even for this density, the adaptive MCMC algorithm finds optimal θ so that more samplings are done near $V = v$. This twist also samples more near the center of the distribution. This sounds like a waste of samples, but this part of sampling hels stablizing the fluctuation of the normalization constant. Therefore, we can make the analogous connection of this part of the distribution to the first run of the method from the previous paragraph. The peak at $V = v$ can be compared to the conditional probability for the second run of the method from the previous paragraph. Our experiments have shown that this twist has the similar effect as the method described in the previous paragraph in reducing variance for the normalization constants. One could also use the Robbins-Monro algorithm to optimize this parameter C at the same time. We have not tried this approach yet.

Chapter 6

Convergence Analysis and Theory for Adaptive MCMC

6.1 Stochastic-Approximation (SA)

6.1.1 General Robbins-Monro (RM) Algorithm

Finding an optimal twisted density is the key to the successful importance sampling. For MCMC or Dynamic Monte Carlo cases, it is very difficult, if not impossible to find such density analytically. However, the Robbins-Monro algorithm from stochastic approximation theory allows us to find such a density through MCMC simulation. The Robbins-Monro algorithm is a stochastic approximation iteration scheme generally taking of the form:

$$\theta_{n+1} = \theta_n + \gamma_n H(\theta_n, X_n) \tag{6.1}$$

where γ_n is a slowly decreasing sequence (to be defined later), and X_n is a d-dimensional random numbers (either i.i.d., or Markov under some restrictive conditions). By the above iteration, it is clear that θ_n becomes random sequence. The iteration (6.1) can be thought as the stochastic version of the Newton's method in deterministic root finding routine. Let us define the following:

$$h(\theta) \equiv E_\theta(H(\theta, x)) \tag{6.2}$$

where θ is a m-dimensional parameter family taking values in non-negative reals R^{m+} (our case $m = 1$). The expectation is taken over $x \in R^d$ under a density parameterized by θ. Note that the parameter family can be defined in general real space, but we shall limit our case to non-negative reals for our theoretical argument. In general, one wish to find θ^* such that :

$$h(\theta^*) \equiv E_{\theta^*}(H(\theta^*, x)) = 0 \tag{6.3}$$

For our Robbins-Monro process (5.16), $H(\theta, x) = sign(V - v)$, and $h(\theta)$ is defined as:

$$h(\theta) = E_\theta(H(\theta, x)) = \frac{\int sign(V - v) f e^{-\theta \cdot \max(V,v)} dx}{\int f e^{-\theta \cdot \max(V,v)} dx} \tag{6.4}$$

Let us further define the following decomposition:

$$H(\theta_n, X_n) - h(\theta) = \xi_n + e_n = \rho_n \tag{6.5}$$

where ρ_n is the total random perturbation term for the step n, ξ_n is the Martingale increment part, and e_n is the small remainder (which is non-zero only for Markov case). The above representation is the Doob's decomposition of a general non-martingale sequence. Then the expression (6.1) can be re-written as:

$$\theta_{n+1} = \theta_n + \gamma_n h(\theta_n) + \gamma_n \rho_n \tag{6.6}$$

The above expressions are very useful from theoretical view points, and we shall use these notations throughout this paper.

6.1.2 Lyapunov Condition for RM Algorithm

In order for this Robbins-Monro process to work (meaning assuring its convergence and the target density (5.12) having invariant distribution which samples more in the VAR region), several technical conditions need to be satisfied. They are the existence of the Lyapunov function (the algorithm stays within the correct region), the A-Stability (stability of the algorithm), and most importantly the ergodicity of the final target density (5.12). The last item allows us to treat the MCMC simulation just like i.i.d. case in that the strong law of large numbers and the Central Limit Theorem still hold. Let us first define the Lyapunov Condition and the Lyapunov function W.

Definition 6.1.1 (Lyapunov Condition)
Given a continuous function (or field) $h(\theta)$ defined in a domain $O \subset R^{m+}$, there exist a function $W(\theta)$ defined in the domain O such that

(1) $W(\theta)$ is non-negative and C^1 (i.e. continuously differentiable)

(2) $W(\theta) \to \infty$ if $\theta \to \partial O$ or $\theta \to \infty$

(3) $\forall \theta \in O$, $\langle \nabla W(\theta), h(\theta) \rangle \leq 0$

(4) There is a compact set $K \subset O$ such that $\langle \nabla W(\theta), h(\theta) \rangle < 0$ if $\theta \notin K$

(5) $W(B)$ has an empty interior if $B = \{\theta; \langle \nabla W(\theta), h(\theta) \rangle = 0\}$

The function W in the above condition is called the Lyapunov function. The set B is often called a set of 'critical values'. The Lyapunov function W tells us that the function h is attractive enough even when θ is too far or close to ∂O. The condition (5) in the above Lyapunov condition is often validated by showing that W is $C^d, d \geq 2$ (all partial derivatives of order up to $d (d \geq 2)$ for the function W are continuous). It is guaranteed by the following theorem ([14]) from differential geometry which we will state without proof.

Thorem 6.1.2 (Sard's Theorem)
If function W and the set of critical values B are given as in the Lyapunov condition, then $W(B)$ has an empty interior if W is $C^d, d \geq 2$.

Let us see if there is a good Lyapunov function W for our case. The existence of W is one of several requirements for proving that our Robbins-Monro algorithm gives almost sure convergence for θ_n to some limit $\theta^* \in K \subset O$. This is also important for showing that π_{θ^*} as given in (5.12) is the stationary distribution for our MCMC where θ^* is given by the above algorithm. For this purpose, we would like to propose the following theorem.

Thorem 6.1.3 (Existence of a Lyapunov Function W for (5.16))
Given the Robbins-Monro process in (5.16) and the proposed target density as in (5.12), let us assume that the portfolio function V is defined in R^d, and $\theta \in R^+$. Let us define function $h(\theta)$ as in (6.2). Let us assume that the quantile p for the VAR is strictlty less than 0.5 ($p < 0.5$). Then there exists an unique θ^* such that $h(\theta^*) = E_{\theta^*}(H(\theta^*, x)) = 0$, and there exists a compact set $K \subset R^+$ such that $\theta^* \in K$. Furthermore, there exists a Lyapunov function W for h given by:

$$W(\theta) = e^{(\theta - \theta^*)^{2 \cdot m}} \tag{6.7}$$

where m is any positive integer.

Note that θ^* is not known in advance (otherwise the Robbins-Monro algorithm is not necessary), so the function W above is created purely for our theoretical purpose.

Proof: We shall prove the existence and the uniqueness of θ^* first. If $\theta = 0$, then the density (5.12) reduces itself to a simple d-dimensional correlated lognormal density. By definition, VAR v is located at the tail end of the distribution for the portfolio function V. Therefore, when $\theta = 0$, more samples result in $(V - v) \geq 0$. Computing h gives $h = -(p) + (1 - p) = 1 - 2p > 0$ since $p \leq 0.5$ by our hypothesis. Therefore $h(0) > 0$ (Actually, this corresponds to p-th quantile). (Note that if $p = 0.5$, then $h = 0$ which corresponds to $\theta = 0$. So no

Robbins-Monro algorithm is necessary). Also by the definition of VAR, there are instances of $V - v = 0$ (otherwise, v would not be VAR).

The density (5.12) has the characteristic that for positive θ, the density will be larger for the region where $(V \leq v)$, and smaller for the other region. As θ gets increased, the density for the region corresponding to $(V \leq v)$ will be exponentially larger, and the other region will be exponentially smaller. Thus letting $\theta \to \infty$ result in $h(\theta) \to -1$ since for this case $(V \leq v)$ with probability 1. It is also clear that h is a strictly monotone decreasing function of θ going from $1 - 2p > 0$ to -1 (changing sign). Since h is a continuous function of θ, there exists exactly one root θ^* satisfying $h(\theta^*) = 0$ by standard calculus result. Thus the existence and uniqueness is proved.

Also following the same argument from above, it is clear that we can pick any arbitrary small $-\delta > 0$ such that for this δ, there exists an unique θ^{**} such that $h(\theta^{**}) = -\delta$. Thus, there exists a compact set $K = [0, \theta^{**}]$ for this Robbins-Monro process.

Next we shall prove that our choice of W is indeed the Lyapunov function for h. The function h is continuous (Actually differentiable too with respect to θ). The function W is clearly continuous and non-negative. As a matter of fact it is C^∞. Thus condition (1) is satisfied. Condition (2) is OK as $\lim_{\theta \to \infty} W(\theta) = \infty$. Next consider $h(\theta) \equiv E_{\theta^*}(H(\theta, x)) = \frac{\int sign(V-v)\pi_\theta}{\int \pi_\theta}$. If $\theta - \theta^* \geq 0$, then larger θ results in reducing the density for the region $(V - v) \geq 0$, and increase the density for the region $(V - v) \leq 0$ This will results in more instances of $(V - v) \leq 0$. Therefore $h(\theta) \leq 0$. Similarly, if $\theta - \theta^* \leq 0$, then smaller θ results in increasing the density for the region $(V - v) \geq 0$, and reducing the density for the region $(V - v) \leq 0$ Therefore $h(\theta) \geq 0$. Since $\frac{\partial W}{\partial \theta} = 2 \cdot m(\theta - \theta^*)^{2 \cdot m - 1} W$, we have $\frac{\partial W}{\partial \theta} \cdot h \leq 0, \forall \theta \in R^+$. Thus the condition (3) is satisfied. We have the compact set $B = \{\theta^*\}$ (consists of only one point) from the above reasoning. This compact set B is the set of critical values. The above arguments show the strict inequality for R^+ outside B. Thus condition (4) follows. Finally, for this B, we have $B = \{\theta; \langle \nabla W(\theta), h(\theta) \rangle = 0\}$. Since W is C^∞, by Sards Theorem, $W(B)$ has an empty interior. One can also see without using Sards Theorem that since it consists of only one point, $W(B)$ has an empty interior.

QED

Thus the Robbins-Monro algorithm will keep increasing itself until $E_{\theta^*}(H(\theta^*, x)) = 0$ if initially started with small positive θ_0, or keep decreasing if initially started with large positive θ_0. From the above argument, we can safely define our new set $O \subset R^+$ as the compact set K we found above. Therefore, we shall use the fact that θ for our Robbins-Monro is contained in some compact set K.

6.1.3 A-Stability for RM Algorithm

So far, we argued that θ tends to change monotonically based on argument using the expectation h of H. However, during the Robbins-Monro process, θ fluctuates because it is stochastic (whether i.i.d. or Markov). Therefore, the notion of stability is important to practically control the fluctuation of θ_n for successful convergence. The stability is a very important concept for the iteration like this. A condition called the 'A-stability' gives the stability for the above iteration.

Definition 6.1.4 (A-Stability Condition)
The algorithm (6.1) is A-stable if

(1) θ_n *remains in a compact subset of* O

(2) $\lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i \xi_i$ *exists*

(3) $\lim_{n\to\infty} |e_n| = 0$

Definition 6.1.5 (Slowly Decreasing Condition)
The algorithm (6.1) is slowly decreasing if γ_n *is a non-negative sequence satisfying*

(1) $\sum_{i=1}^{\infty} \gamma_i = \infty$

(2) $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$

As one can see from the Robbins-Monro algorithm that there are deterministic part, and the stochastic part of the algorithm. The Lyapunov condition, the slowly decreasing condition on γ_n, and the A-Stability of the algorithm aid the proof for the convergence for the ODE part of the algorithm. Let us state a general theorem, and simple proof based on (Delyon [14]). The proof explains the importance of the Slowing varying condition very well. Thus, it is worth exploring the proof here.

Thorem 6.1.6 (General Convergence Theorem: ODE Part)
Let us assume that the algorithm (6.1) is A-stable, slowly decreasing, and the function h of the algorithm satisfies the Lyapunov condition. Then the distance from θ^* *from the set of critical values* B *goes to zero, i.e.* $d(\theta_n, B) \to 0$. *If the set* B *is finite, then* θ_n *converges to some point of* B.

Proof: Let us define $\hat{\theta}_{n+1}$ as:

$$\hat{\theta}_{n+1} = \theta_{n+1} + \sum_{i=n+1}^{\infty} \gamma_i \xi_i = \theta_{n+1} - \delta_{n+1} \tag{6.8}$$

where $\delta_{n+1} = -\sum_{i=n+1}^{\infty} \gamma_i \xi_i$. Then we can express $\hat{\theta_{n+1}}$ as:

$$\begin{aligned}
\hat{\theta}_{n+1} &= \theta_n + \gamma_n h(\theta_n) + \gamma(\xi_n + e_n) + \sum_{i=n+1}^{\infty} \gamma_i \xi_i \\
&= \theta_n + \sum_{i=n}^{\infty} \gamma_i \xi_i - \gamma_n \xi_n + \gamma_n h(\theta_n) + \gamma(\xi_n + e_n) \\
&= \hat{\theta}_n + \gamma_n h(\hat{\theta}_n + \delta_n) + \gamma_n e_n \\
&= \hat{\theta}_n + \gamma_n h(\hat{\theta}_n) + \gamma_n \bar{e}_n \qquad (6.9)
\end{aligned}$$

where $\hat{e}_n = e_n + h(\hat{\theta}_n + \delta_n) - h(\hat{\theta}_n)$. Since by (2) of the A-stability condition, $\delta_n \to 0$. Therefore $\hat{\theta}_n$ remains in some compact set K_0 since θ_n does so by (1) of the A-Stability condition. (3) of the A-Stability condition guarantees that $e_n \to 0$. Furthermore, the function h is uniformly continuous on compact set. Therefore $\hat{e}_n \to 0$. Since W is C^1, the first order Taylor expansion of W gives:

$$\begin{aligned}
W(\hat{\theta}_{n+1}) &= W(\hat{\theta}_n) + \gamma_n \langle \nabla W(\theta_M), h(\hat{\theta}_n + \hat{e}_n) \rangle \\
&= W(\hat{\theta}_n) + \gamma_n \langle \nabla W(\theta_n), h(\hat{\theta}_n) \rangle + \gamma_n \bar{e}_n \qquad (6.10)
\end{aligned}$$

where θ_M is a point between θ_n and θ_{n+1}, and $\bar{e}_n$ is given by:

$$\bar{e}_n = \langle \nabla W(\theta_M) - \nabla W(\theta_n), h(\hat{\theta}_n) + \hat{e}_n \rangle + \langle \nabla W(\theta_n), \hat{e}_n \rangle \qquad (6.11)$$

Notice that $\hat{\theta}_{n+1} - \hat{\theta}_n \to 0$ as $n \to \infty$ because h is bounded in a compact set and $\gamma_n \to 0$. Therefore $\nabla W(\theta_M) - \nabla W(\theta_n) \to 0$ as $n \to \infty$ since ∇W is continuous and θ_M is a point between θ_n and θ_{n+1}. Similarly ∇W is bounded on a compact set. Thus $\bar{e}_n \to 0$ as $n \to \infty$. Take any open set Q of $B \cap K_0$. Since $\langle \nabla W(\theta_n), h(\hat{\theta}_n) \rangle$ is continuous on a compact set, it will attain its maximum on the set. Therefore there exist $\varepsilon > 0$ such that for θ_n in the compact set $K_0 \setminus Q$, we have:

$$\langle \nabla W(\theta_n), h(\hat{\theta}_n) \rangle \leq -2\varepsilon \qquad (6.12)$$

Since $\langle \nabla W, h \rangle \leq 0$ in K_0 but equals to 0 in B (by the Lyapunov condition), $\langle \nabla W, h \rangle$ will be closer to 0 in the small neighborhood Q by its continuity. Thus there exist a constant $C > \varepsilon > 0$ such that for a sufficiently large n, we have:

$$W(\hat{\theta}_{n+1}) \leq W(\hat{\theta}_n) - \gamma_n \varepsilon + \gamma_n C I_Q(\hat{\theta}_n) \qquad (6.13)$$

where I_Q is an indicator function.

Define the neighborhood of $W(B \cap K_0)$, Q_a as $Q_a = \{x \in R : d(x, W(B \cap K_0)) < a)\}$ where $a > 0$. Q_a is basically a union of interval of length $2a$. Let us define $Q = W^{-1}(Q_a)$ (the inverse image of Q_a). Then we for this Q_a, we have:

$$W(\hat{\theta}_{n+1}) \leq W(\hat{\theta}_n) - \gamma_n \varepsilon + \gamma_n C I_{Q_a}(\hat{\theta}_n) \qquad (6.14)$$

Notice that when $W(\hat{\theta}_n)$ leaves the set Q_a ($W(\hat{\theta}_n) \notin Q_a$), the new $W(\hat{\theta}_{n+1})$ decreases by $\gamma_n \varepsilon$. Since $-\sum \gamma_n = -\infty$, $W(\hat{\theta}_n)$ will eventually reach another interval of Q_a with smaller $W(\hat{\theta}_{n+1})$. On the other hand, when $W(\hat{\theta}_n) \in Q_a$, it will increase by $\gamma_n(C-\varepsilon)$. Thus $d(W(\hat{\theta}_n), Q_a) \to 0$ as $n \to \infty$. Since $W(B \cap K_0)$ is compact, the set of limit points of $W(\hat{\theta}_n)$ is contained in $W(B \cap K_0)$. Since $W(\hat{\theta}_{n+1}) - W(\hat{\theta}_n) \to 0$, this limit points is an interval. Moreover, $W(B)$ has an empty interior by the Lyapunov condition. Thus $W(\hat{\theta}_n)$ converges to a point of $W(B)$.

Next, take again any arbitrary open neighborhood Q of B. Since $W(\hat{\theta}_n)$ converges, for any $\delta > 0$, there exists $n(\delta)$ such that:

$$|W(\hat{\theta}_n) - W(\hat{\theta}_m)| < \delta, \quad n(\delta) < n < m \tag{6.15}$$

Since by (6.14), $\gamma_n\varepsilon + \gamma_n C I_{Q_a}(\hat{\theta}_n) \leq W(\hat{\theta}_n) - W(\hat{\theta}_{n+1})$, we have the following by the telescoping sum:

$$\varepsilon \sum_{i=n+1}^{m} \gamma_i - C \sum_{i=n+1}^{m} \gamma_i I_{Q_a}(\hat{\theta}_i) < \delta, \quad n(\delta) < n < m \tag{6.16}$$

Let us take any m such that $\varepsilon(\sum_{i=n+1}^{m} \gamma_i) > \delta$ (This is possible since $\sum \gamma = \infty$). Since the first term of the expression (6.16), there exists $k \in [n, m]$ such that $\hat{\theta}_k \in Q$. Since by (6.16), $\hat{\theta}_{n+1} - \hat{\theta}_n = \gamma_n h(\hat{\theta}_n) + \gamma_n \bar{e}_n = O(\gamma_n)$, we have: $\hat{\theta}_k - \hat{\theta}_n = O(\delta/\varepsilon)$. Thus $\hat{\theta}_n$ is close to Q by a distance of $O(\delta/\varepsilon)$. Since δ is arbitrary, letting $\delta \to 0$ gives $d(\hat{\theta}_n, Q) \to 0$. Since Q is arbitrary (or a is arbitrary), $\hat{\theta}_n \to B$.

QED

It is clear from the proof that the slowly decreasing condition is very important in guaranteeing that θ is getting pulled to a critical point θ^*. Too fast or too slow decrease of γ_n does not help the convergece. The Lyapunov Condition makes the Taylor expansion of W possible and helps this attractor to the critical set by maintaining the inequality (6.14) throughout the algorithm. Finally the notion of the A-Stability is used to make the set $W(B \cap K_0)$ being compact which guarantees that the convergence of the subsequence in the set with limit $W(\theta*)$ which is the element of the set. The above theorem is nice, but the notion of the A-Stability is hard to check in general. For our Robbins-Monro algorithm, we have proven that our θ remains in some compact set $K \in R^+$. It will be shown later that for our MCMC chain, A-stability can be shown. However, in general, it is very difficult to check if θ does not wander into infinity. In order to counter this problem there is a nice algorithm due to Chen ([10]) by performing projections to reduce some problems into bounded cases under the recurrent conditions. The details will be given below. The idea is to introduce the notion of recurrence without spoiling the convergence. We have tested this algorithm, and appear to help our algorithm for some portfolios. So we shall describe this algorithm.

6.1.4 Stability by Projection: Chen's Algorithm

There is a systematic way to create an A-stability based on a simple algorithm. It is based on performing 'projection' finite number of times without affecting the result of the algorithm. In this section below, we shall describe the basic notion of recurrence and some weaker conditions that guarantees the A-Stability when the Chen's algorithm is implemented. Let us first describe the Chen's algorithm.

Definition 6.1.7 (Chen's Algorithm)
Choose any increasing sequence of compact sets $(K_n)_{n\geq 0}$ such that $\cup_n K_n = O$. The following modified Robbins-Monro algorithm is called the Chen's algorithm if the iteration is given by:

$$\hat{\theta}_{n+1} = \theta_n + \gamma_n H(\theta_n, X_n) \tag{6.17}$$

where θ_{n+1} is updated by the following procedure

$$\theta_{n+1} = \hat{\theta}_{n+1} \text{ and set } s(n) = s(n-1), \text{ if } \hat{\theta}_{n+1} \in K_{s(n-1)}$$

$$\theta_{n+1} = \bar{\theta}_{n+1} \in K_0 \text{ and set } s(n) = s(n-1)+1, \text{ if } \hat{\theta}_{n+1} \notin K_{s(n-1)}$$

This has the effect of pulling back the θ_{n+1} to the point in the initial compact set K_0 and increasing the size of new compact set $K_{s(n+1)}$. This is also nice theoretically, and its implementation is quite easy. With this algorithm being implemented, we have the following lemma which we will state without proof from Delyon [14].

Lemma 6.1.8 (Projection Lemma)
Given the Chen's algorithm in (6.17), and the compact set K_n as given in the algorithm. Let us assume that the Lyapunov condition is also satisfied. Let $\tau(n)$ be the last projection time before n ($\tau(n) < n$). For some large m, if the following condition is satisfied:

$$\lim_{n\to\infty} I(\theta_{\tau(n)}, ...\theta_n \in K_m)\{|e_{n+1}| + |\sum_{i=\tau(n)+1}^{n+1} \gamma_i \xi_i|\} = 0 \tag{6.18}$$

where I is an indicator function. Then the sequence θ_n is contained in come compact set K_m, and projections are done only finite times.

When checking the A-stability using Chen's algorithm, the key condition is to satisfy the expression (6.18). Let us define a sequence of compact set as $K_m \equiv \{\theta : W(\theta) \leq m\}$. Then it is clear that the following theorem satisfies the projection lemma since the conditions (6.19) and (6.20) implies the expression (6.18). Therefore, we have the following theorem which is easier to apply later on.

Thorem 6.1.9 (Projection Theorem)
Given the Chen's algorithm in (6.17), and the compact set K_n as given in the algorithm. Let us assume that the Lyapunov condition is also satisfied. If for any $m \in N$, the following condition is satisfied:

$$\lim_{m\to\infty} \sum_{i=1}^{m} \gamma_i \xi_i I_{W(\theta_{i-1})\leq m} < \infty \tag{6.19}$$

$$\lim_{n\to\infty} |e_n| I_{W(\theta_{n-1})\leq m} < \infty \tag{6.20}$$

Then the algorithm is A-stable, and the number of projections are finite.

The idea of this theorem is that under the Chen's method, if we have the negligible (decreasing to 0 as $n \to \infty$) tail series for both $\sum_{i=n}^{\infty} \gamma_i \xi_i$ and e_n, then we have the A-stability for the algorithm. Notice that $\sum_{i=n}^{\infty} \gamma_i \xi_i$ is a martingale (difference) part, and e_n is an non-martingale noise. Since ξ_i is a martingale difference, we have the following relation: $E[\sum_{i=1}^{n+1} \gamma_i \xi_i | F_n] = \sum_{i=1}^{n} \gamma_i \xi_i \quad \forall n$ where F_n is a standard σ-field. To show the negligibility of the martingale part, the following theorem is often useful. We shall quote this theorem without proof. The proof can be found in various probability text books (Chung[34]).

Thorem 6.1.10 (Martingale Convergence Theorem)
If F_n is an increasing sequence of σ-fields, and Z_n is a martingale If , for some $1 \leq r \leq 2$, the following martingale difference $Y_n = Z_n - Z_{n-1}$ satisfies the condition:

$$\sum_{n=1}^{\infty} E[|Y_n|^r | F_{n-1}] < \infty \quad a.s \tag{6.21}$$

Then $Z_n = \sum_{i=1}^{n} Y_i$ converges almost surely.

In the next section, we shall finally describe the general Robbins-Monro Theorem for i.i.d. chain.

6.1.5 General Robbins-Monro Theorem

The Robbins-Monro algorithm satisfying the Lyapunov condition, the A-Stability, and the Slowly decreasing condition is guaranteed to assure almost sure convergence of $\theta_n \to \theta^*$ if H is bounded and the chain is i.i.d. We shall state the general theorem, and provide a very simple proof for it (again following Delyon [14]).

Thorem 6.1.11 (Robbins-Monro Theorem: I.I.D. Case)
Let us assume that the algorithm (6.17) satisfies the following conditions:

(1) The function h of the algorithm satisfies the Lyapunov condition.

(2) The random numbers (X_n) are generated from P_{θ_n} independently from the past (i.i.d.): $P(X_n \in A|X_{n-1}, X_{n-1}, ...X_1, \theta_0) = P_{\theta_n}(A)$.

(3) γ_n satisfies the Slowly decreasing condition.

(4) For any compact set $K \subset O$, $\sup_{\theta \in K} E_\theta(|H(\theta, x)|^2) < \infty$.

Then $\theta_n \to \theta^*$ such that $h(\theta^*) = 0$ almost surely.

Proof: The proof is very simple using the Martingale convergence theorem. Since this is i.i.d., the additive noise is zero ($e_n = 0 \ \ \forall n$). Therefore, we will prove that the martingale part converges almost surely. This will show the A-Stability of the chain almost surely by the Projection Theorem, and therefore the General Convergence Theorem from the previous section shows that $\theta_n \to \theta^*$ where $h(\theta^*) = 0$ almost surely. Let us define $Y_n = \gamma_n \xi_n I_{W(\theta_{n-1}) \le N}$. Then, we have the following relation:

$$\begin{aligned}
\sum E[|Y_n|^2|F_{n-1}] &= \sum \gamma_n^2 I_{W(\theta_{n-1}) \le N} E_{\theta_{n-1}}[|\xi_n|^2] \\
&= \sum \gamma_n^2 I_{W(\theta_{n-1}) \le N} E_{\theta_{n-1}}[|H(\theta_n, X_n) - h(\theta_n)|^2] \\
&\le \sum \gamma_n^2 \sup_{W(\theta) \le N} E_\theta[|H(\theta, x) - h(\theta)|^2] \\
&\le \sum \gamma_n^2 \cdot C_N < \infty \qquad (6.22)
\end{aligned}$$

The last steps are justified because of the boundedness of $E_\theta(|H(\theta, x)|^2)$ in some compact set, the Lyapunov condition, and finally the Slowly decreasing condition of γ_n ($\sum \gamma_n^2 < \infty$). Therefore, by the Martingale convergence theorem, $Z_n = \sum_{i=1}^n Y_i$ converges almost surely, and the hypothesis for the Projection Theorem is satisfied almost surely.

QED

Notice that for our Robbins-Monro algorithm, the condition (4) of the above theorem is trivially satisfied since $|H| = |sign(V - v)| \le 1$. In the next section, we shall extend the Robbins-Monro theorem to our MCMC chain (Metropolis-within-Gibbs Metropolis), and show a convergence theorem for our adaptive MCMC chain.

6.2 Robbins-Monro Algorithm for Adaptive MCMC

In the previous sections, we have shown that our algorithm satisfies the Lyapunov condition, the Slowly decreasing condition, and the A-Stability of the chain assuming i.i.d. This was done because of our choice of H which is bounded absolutely by 1. For our algorithm using the Metropolis-within-Gibbs algorithm, the non-martingale additive noise is not zero. Therefore, we have to establish the A-Stability for MCMC differently. One concept that can aid us in proving this is called the Poisson Condition (Delyon [14]). This condition is satisfied if the chain is geometrically ergodic (it will be defined later) for a fixed $\theta > 0$, and if the continuity of the Metropolis transition probability operator T_θ with respect to θ is proven. In this section, we shall give series of definition and theorem leading to proving that our adaptive Robbins-Monro indeed assure convergence of θ almost surely.

6.2.1 Poisson Equation

The general theory is quite involved. However, since our chain is bounded from above (meaning both the target density π_θ, the Metropolis transition operator T_θ, and most importantly, θ is contained in some compact set $K \subset R^+$, the argument wil be a little simpler. Let us first define $T_\theta g$ to mean an integral operator T_θ acting on positive function g. Therefore, $T_\theta g \equiv \int_X g(x) T_\theta(x, dy) = E_\theta[g(x)]$. Let us explain our transition operator T_θ more concretely. The expressions are quite involved for some representations. However; the detail will be helpful in understanding some of advance concept from the ergodic theory. In our algorithm we can use either Sequential or Random Scan Metropolis-within-Gibbs method. Let us start with the Sequential Metropolis-within-Gibbs operator. Our componentwise Metropolis-within-Gibbs algorithm admits the following T_θ with respect to acceptance and rejection regions in the direction of i-th coordinate:

$$\begin{aligned} T^i_{\theta_n}(x_i) &= \int_{A_i(x_i)} q(|x_i - y_i|) dy_i \\ &+ \int_{R_i(x_i)} q(|x_i - y_i|) a_{\theta_n}(x_i|y_i) dy_i \\ &\quad + I_{B_i(x_i)} \int_{R_i(x_i)} q(|x_i - y_i|)(1 - a_{\theta_n}(x_i|y_i)) dy_i \\ &= \int_{B_i} a_{\theta_n}(x_i|y) q(|x_i - y_i|) dy_i \\ &\quad + I_{B_i(x_i)} \int_{R_i(x_i)} q(|x_i - y_i|)(1 - a_{\theta_n}(x_i|y_i)) dy_i \qquad (6.23) \end{aligned}$$

where

$$A_i(x_i) = \{y_i \in B_i : a_{\theta_n}(x_i|y_i) = \frac{\pi_\theta(y_i)}{\pi_\theta(x_i)} \geq 1\}.$$
$$R_i(x_i) = \{y_i \in B_i : a_{\theta_n}(x_i|y_i) = \frac{\pi_\theta(y_i)}{\pi_\theta(x_i)} < 1\}. \tag{6.24}$$

Therefore, $T^i_{\theta_n} g$ cab be written as:

$$\begin{aligned} T^i_{\theta_n} g(x) &= \int_{B_i} a_{\theta_n}(x_i|y_i) q(|x_i - y_i|) g(y) dy_i \\ &\quad + g(x) \int_{R_i(x_i)} q(|x_i - y_i|)(1 - a_{\theta_n}(x_i|y_i)) dy_i \end{aligned} \tag{6.25}$$

where g is a positive function defined in some set $B = (B_1, B_2, ..., B_d) \subset R^d$. For the Sequential Metropolis-within-Gibbs algorithm, $T^j_{\theta_n} T^i_{\theta_n} g$ can be written as:

$$\begin{aligned} T^j_{\theta_n} T^i_{\theta_n} g(x) &= \int_{B_j} a_{\theta_n}(x_j|y_j) q(|x_j - y_j|) T^i_{\theta_n} g(y) dy_j \\ &\quad + T^i_{\theta_n} g(x) \int_{R_j(x_j)} q(|x_j - y_j|)(1 - a_{\theta_n}(x_j|y_j)) dy_j \end{aligned} \tag{6.26}$$

Continuing this way, when one need to find $\prod_{i=1}^d T^i_{\theta_n} g$ for the entire sequential sweep of the dimension, we have:

$$\begin{aligned} \prod_{i=1}^d T^i_{\theta_n} g(x) &= \int_{B_d} a_{\theta_n}(x_d|y_d) q(|x_d - y_d|) \prod_{i=1}^{d-1} T^i_{\theta_n} g(y) dy_d \\ &\quad + \prod_{i=1}^{d-1} T^i_{\theta_n} g(x) \int_{R_d(x_d)} q(|x_d - y_d|)(1 - a_{\theta_n}(x_d|y_d)) dy_d \end{aligned} \tag{6.27}$$

Therefore, we can see that the sequential Metropolis-within-Gibbs operator is characterized by product operator $T_{\theta_n} = \prod_{i=1}^d T^i_{\theta_n}$. This expression is quite complex algebraically when the dimension becomes larger.

The Random Scan Metropolis-within-Gibbs on the other hand is quite easy. Since by its definition, the sweeps are done randomly, not conditionally on the previous chain of the dimension being finished. Therefore any product of operators corresponding to each dimension is run with probability $\frac{1}{d}$. Thus, for the Random Scan Metropolis-within-Gibbs, we have the following expressions for the total operator for the entire chain.

$$T_{\theta_n} g(x) = \frac{1}{d} \sum_{i=1}^d \int_{B_i} a_{\theta_n}(x_i|y_i) q(|x_i - y_i|) g(y) dy_i$$

$$+g(x)\frac{1}{d}\sum_{i=1}^{d}\int_{R_i(x_i)} q(|x_i-y_i|)(1-a_{\theta_n}(x_i|y_i))dy_i \tag{6.28}$$

The Random Scan Metropolis-within-Gibbs Operator is characterized by the summation (or the average) operator $T_{\theta_n} = \frac{1}{d}\sum_{i=1}^{d} T^i_{\theta_n}$.

We explained operators associated with the Metropolis-within-Gibbs Operators. In our algorithm, the whole key lies in the expression of T_θ which provides us with a excellent theoretical support. This is done because of our choice of the family of target densities $fe^{-\theta max(V,v)}$ which are bounded from above. The key concept that will link our transition operator T_θ to the proof of the general convergence of θ lies in the notion of Poisson Condition (Delyon [14]). We shall define this condition and explain its meaning later:

Definition 6.2.1 (Poisson Condition)
Let us given the transition probability operator T_θ as given as before. Then we say that the transition operator T_θ satisfies the Poisson condition when the following requirements are met:

(1) For every θ, there exists a function $h(\theta)$ such that the following Poisson equation:

$$(I - T_\theta)v_\theta = H(\theta,.) - h(\theta) \tag{6.29}$$

admits a solution v_θ.

(2) For any compact set $K \subset O$, the following is satisfied:

$$\sup_{\theta\in K, x\in X} |H(\theta,x)| + |v_\theta(x)| < \infty \tag{6.30}$$

(3) There is a continuous function ϕ_K such that $\phi_K(0) = 0$ and for any θ, $\hat{\theta}$, we have:

$$\sup_{x\in X} |T_\theta v_\theta(x) - T_{\hat{\theta}} v_{\hat{\theta}}(x)| < \phi_K(|\theta - \hat{\theta}|) \tag{6.31}$$

These conditions do not look helpful at a first glance. Furthermore, it is very difficult to prove that our chain satisfies these conditions. There is another notion called the 'Geometric Ergodicity'. There is a theorem by (Delyon [14]) which guarantees the Poisson Condition when basically two requirements are met - the Geometric Ergodicty of the chain for a fixed θ, and the continuity of the transition operator T_θ with respect to its parameter θ. In the next sections, we shall prove that our MCMC algorithm is geometric ergodic. In the following section after that, we shall prove the continuity of our chain, and present a theorem by Delyon ([14]) which will indeed conclude that our chain satisfies the

Poisson Condition.

For a present moment, we shall assume that our chain is obtained via the Random Scan Metropolis-within-Gibbs algorithm (Average Operator). The convergence proof for the Sequential Scan Metropolis-within-Gibbs algorithm is hard to show because of the algebraic complexities involved in proving its geometric ergodicity. Therefore, we shall not take up this case in this paper.

6.2.2 Geometric Ergodicity

Let us define the notion of geometric ergodicity in general. Since this paper is not about the ergodic theory in general, we shall be brief in explaining definitions and theorems concerning the ergodic theory. Before beginning with this concept, let us define the V-normed space as the space of continuous functions C_V with the V-norm defined as:

$$C_V = \{g \text{ continuous finction defined in } R^d : \sup_{x \in X \subset R^d} \frac{|g(x)|}{V(x)} < \infty\} \tag{6.32}$$

where V is a positive function in R^d such that $V \geq 1$. The space is equipped with the V-norm : $||g||_V \equiv \sup_{x \in X} \frac{|g(x)|}{V(x)}$ where $g \in C_V$. Furthermore its operator norm can be defined as: $||T_\theta||_V \equiv \sup_{g \in C_V} \frac{||Tg||_V}{||g||_V}$. The V-normed space is a very special norm giving an extra degree of freedom for allowing any V to be a candidate as long as $V \geq 1$. Let us define the notion of the V-uniform ergodicity.

Definition 6.2.2 (V-Uniform Ergodicity)
Let us be given a transition probability operator T for a given Markov chain. Assume that transition T is aperiodic and irreducible with an invariant distribution π, Then the Markov chain is said to be V-Uniform Ergodic if there exist a positive function $V \to [1, \infty]$, a constant $C > 0$, and a constant $\rho < 1$ such that

$$||T^n(x, \cdot) - \pi||_V \leq C \cdot \rho^n \cdot V(x) \quad n = 1, 2, ... \tag{6.33}$$

for π-almost all $x \in X$ where $||||_V$ is a V-norm defined above.

Let us further define the geometric ergodicity as:

Definition 6.2.3 (Geometric Ergodicity)
Let us the transition operator T as given above. Assume that transition T is

aperiodic and irreducible with an invariant distribution π, Then the Markov chain is said to be Geometric Ergodic if there exist a constant $\rho < 1$ and a positive constant depending on x such that $C_x < \infty$ for fixed $x \in X$ satisfying:

$$\|T^n(x,\cdot) - \pi\| \leq C_x \cdot \rho^n \quad n = 1, 2, ... \tag{6.34}$$

for π-almost all $x \in X$ where $\|T\| = \frac{1}{2} \sup_{|g| \leq 1} |Tg|$.

$\|\|$ above is often called the total variation norm. The geometric ergodicity basically says that the convergence rate of a chain to the steady state distribution π is achieved at a geometric rate. The V-uniform ergodicity is basically similar to the geometric ergodicity except for the dependency on the function V. It is proved by Boberts & Rosenthal ([61]) that if the chain is aperiodic and irreducible, then the geometric ergodicity and the V-uniform ergodicity are equivalent. Therefore, we can focus on proving that our chain is V-uniform ergodic by picking a right choice of V. The notion of the geometric ergodicity is nice, but hard to prove in general. Typically, the curvature of the target density has to be checked. This involves that the maximum curvature for the density goes to infinity as $|x| \to \infty$ where $|\cdot|$ is a Euclidean norm in R^d. This is not easy to check in general. However; Fort, Moulines, Roberts and Rosenthal (Henceforth, 'FMRR', [27]) proposed a set of conditions for a target density for the Random Scan Metropolis-within-Gibbs algorithm without the knowledge of conditions on the geometry of the contour manifold of the density in order to verify the geometric ergodicity. We shall state their conditions, and several theorems without proofs. In this paper, we shall define the new condition, the 'FMRR Geometric Ergodicity Condition'. Then we will prove that our target density π_θ does indeed satisfy this condition for all $\theta \geq 0$.

Definition 6.2.4 (FMRR Geometric Ergodicity Condition)
Let us be given the target probability measure π defined in R^d with its Borel σ-field $B(R^d)$ equipped with density π_θ as in (5.12). Let us define the usual Lebesgue measure on R^d to be μ_d, and $(e_1, e_2, ...e_d)$ to be the coordinate unit vectors. Let us define $(q_i)_{(i=1,2...d)}$ to be a family of symmetric random-walk Metropolis density as defined previously. Then we shall say that the target density satisfy the 'FMRR Geometric Ergodicity Condition' if the followings are satisfied:

(1) The target distribution π is absolutely continuous with respect to μ_d, and its density π_θ is positive and continuous in R^d.

(2) The family of random-walk densities (q_i) satisfies that there exists constants $\varepsilon_{i\,i=1,2,...d} > 0$, and $\delta_{i\,i=1,2,...d} < \infty$ such that if $|x| \leq \delta_i$, then $q_i \geq \varepsilon_i$.

(3) There is $0 \leq \delta < \Delta \leq +\infty$ so that $\rho = \inf_{1 \leq i \leq d} \int_\delta^\Delta q_i(y)dy > 0$, and for any sequence $x \equiv x^j$ with $\lim_{j\to\infty} |x^j| = +\infty$, there exists a subsequence

$\hat{x} \equiv \{\hat{x}^j\}$ such that for some $i \in 1, 2, ...d$ and all $y \in [\delta, \Delta]$, the followings are satisfied.

$$\lim_{j\to\infty} \frac{\pi_\theta(\hat{x}^j)}{\pi_\theta(\hat{x}^j - sign(\hat{x}^j) y \cdot e_i)} = 0 \tag{6.35}$$

and

$$\lim_{j\to\infty} \frac{\pi_\theta(\hat{x}^j + sign(\hat{x}^j) y \cdot e_i)}{\pi_\theta(\hat{x}^j)} = 0 \tag{6.36}$$

Next, we would like to prove that our target density given in (5.12) does satisfy the FMRR Geometric Ergodicity Condition.

Thorem 6.2.5 (Geometric Ergodicity for the Density π_θ in (5.12)) *Let us be given the target probability density π_θ as in (5.12). Let us also assume that the symmetric random-walk Metropolis density is given as in the section (5.3.4). Then for every $\theta \geq 0$, the density π_θ satisfies the FMRR condition.*

Proof: The target density π_θ is explicit written as:

$$\pi_\theta(x) = C \frac{1}{(2\pi)^{d/2} |\Sigma_x|} e^{-\frac{x^T \cdot \Sigma_x^{-1} \cdot x}{2}} e^{-\theta * max(V(x), v)} \tag{6.37}$$

It is clear that our target density is positive and absolutely continuous with respect to the usual Lebesque measure in R^d. Thus the condition (1) is cleared. Our choice of random-walk Metropolis proposal density is a symmetric gaussiasn density which is positive. Thus the condition (2) is also satisfied.

Next for a given sequence of $(x^j) = (x_1^j, x_2^j, ...x_d^j)$ such that $|x^j| \to \infty$, define the new sequence $\hat{x}^j$ by $x^j = C^T \cdot \hat{x}^j$ where C is obtained by the Cholesky decomposition so that $C^T \cdot C = \Sigma_x$. This is basically diagonalizing (or rotating toward principal component) the matrix expression in (6.37). Thus for this $\hat{x}^j$, we have the following expression:

$$\begin{aligned} \pi_\theta(\hat{x}^j) &= C \frac{1}{(2\pi)^{d/2} |\Sigma_x|} e^{(-\sum_{i=1}^d \frac{(\hat{x}_i^j)^2}{2})} e^{-\theta * max(V(\hat{x}^j), v)} \\ &= \hat{C} e^{-\sum_{i=1}^d \frac{(\hat{x}_i^j)^2}{2}} g(\hat{x}^j) \end{aligned} \tag{6.38}$$

where $g(x) = exp^{(-\theta * max(V(x), v))}$. Pick any positive $y \in R$ such that $0 \leq \delta < y < \Delta \leq +\infty$. Since $|x^j| = \sqrt{\sum_{i=1}^d (x_i^j)^2} \to \infty$, at least one of the coordinate element (x_m^j) must satisfy $|x_m^j| \to \infty$. Since $\hat{x}^j = (C^T)^{-1} \cdot x^j$, at least one of the coordinate element $(\hat{x}_k^j)$ must satisfy $|\hat{x}_k^j| \to \infty$ as well. Let us define $\hat{y}_k^j = \hat{x}^j - sign(\hat{x}^j) y \cdot e_k$; Then computing $\pi_\theta(\hat{y}_k^j)$ gives:

$$\pi_\theta(\hat{y}_k^j) = \hat{C} e^{-\sum_{i=1}^d \frac{(\hat{x}_i^j)^2}{2} + \frac{(\hat{x}_k^j)^2 - (\hat{x}_k^j - sign(\hat{x}^j) y)^2}{2}} g(\hat{y}_k^j) \tag{6.39}$$

Therefore $\frac{\pi_\theta(\hat{x}^j)}{\pi_\theta(\hat{y}^j_k)}$ gives:

$$\begin{aligned}
\frac{\pi_\theta(\hat{x}^j)}{\pi_\theta(\hat{y}^j_k)} &= \frac{g(\hat{x}^j)}{g(\hat{y}^j_k)}e(\frac{-(\hat{x}^j_k)^2 + (\hat{x}^j_k - sign(\hat{x}^j_k)y)^2}{2}) \\
&= \frac{g(\hat{x}^j)}{g(\hat{y}^j_k)}e(\frac{y^2 - 2sign(\hat{x}^j_k)\hat{x}^j_k y}{2}) \\
&= \frac{g(\hat{x}^j)}{g(\hat{y}^j_k)}e(\frac{y^2 - 2|\hat{x}^j_k|y}{2}) \\
&\leq \bar{C}e(\frac{y^2 - 2|\hat{x}^j_k|y}{2}) \qquad (6.40)
\end{aligned}$$

The last step is justified since $\frac{g(\hat{x}^j)}{g(\hat{y}^j_k)}$ is a strictly positive function bounded from above by our construction of g. Thus letting $j \to \infty$ yields:

$$0 \leq \lim_{j\to\infty} \frac{\pi_\theta(\hat{x}^j)}{\pi_\theta(\hat{y}^j_k)} \quad \leq \quad \bar{C} \lim_{j\to\infty} e(\frac{y^2 - 2|\hat{x}^j_k|y}{2}) = 0 \qquad (6.41)$$

Thus, the condition (6.35) is satisfied. We can follow the same argument for the condition (6.36) using $\bar{y}^j_k = \hat{x}^j + sign(\hat{x}^j)y \cdot e_k$, giving:

$$\pi_\theta(\bar{y}^j_k) = \hat{C}e^{-\sum_{i=1}^{d} \frac{(\hat{x}^j_i)^2}{2} + \frac{(\hat{x}^j_k)^2 - (\hat{x}^j_k + sign(\hat{x}^j)y)^2}{2}} g(\hat{y}^j_k) \qquad (6.42)$$

Thus the similar calculation gives

$$\begin{aligned}
0 \leq \frac{\pi_\theta(\bar{y}^j_k)}{\pi_\theta(\hat{x}^j)} &= \frac{g(\hat{x}^j)}{g(\hat{y}^j_k)}e(\frac{(\hat{x}^j_k)^2 - (\hat{x}^j_k + sign(\hat{x}^j_k)y)^2}{2}) \\
&\leq \bar{C}e(\frac{-y^2 - 2|\hat{x}^j_k|y}{2}) \qquad (6.43)
\end{aligned}$$

Thus taking a limit $j \to \infty$, the condition (6.36) is proved.

QED

So we have shown that out target density π_θ does indeed satisfy the FMRR Geometric Ergodicity Condition. According to paper by FMRR, this is enough to show that the Random Scan Symmetric random-walk Metropolis with the target density π_θ is V-uniformly ergodic, and therefore geometrically ergodic. We shall quote a theorem by FMRR (Theorem 4, [27]) without proof.

Thorem 6.2.6 (Geometric Ergodicity for the Random Scan Metropolis-within-Gibbs)
Let us assume that the Random Scan Metropolis-within-Gibbs Markov chain is

aperiodic and irreducible with invariant distribution π_θ for some $\theta \geq 0$. Let us further assume that the FMRR Geometric Condition is satisfied. Let us denote T_θ as the transition operator for the Random Scan Metropolis-within-Gibbs algorithm given by $T_\theta = \frac{1}{d}\sum_{i=1}^{d} T_\theta^i$. Then there exist a function V given by $V = \pi_\theta^{-s}$ for some $s \in (0,1)$ such that

$$T_\theta V(x) \leq \lambda V(x) + b \cdot I_C(x) \tag{6.44}$$

for some $\lambda \in (0,1)$, $b < \infty$, and for some small set $C \in X$. In particulat, T_θ is V-uniformly ergodic, and therefore geometrically ergodic (Proposition 2.1, [61]).

Thus, we have shown that our Random Scan Metropolis-within-Gibbs is geometrically ergodic for a fixed $\theta \geq 0$. In the next section, we shall prove the continuity of the transition operator with respect to θ to conclude that our chain does indeed satisfy the Poisson Condition.

6.2.3 Continuity with Respect to a Parameter for Robbins-Monro MCMC

Thus far, we have shown that our chain is geometrically ergodic for a fixed θ. In order for our algorithm to satisfy the Poisson condition, we have to ensure that the geometric ergodicity holds continuously with respect to θ. In the next theorem, we shall prove that our transition operator T_θ does indeed satisfy the continuity condition.

Thorem 6.2.7 (θ Continuity Theorem)
Let us assume that the Random Scan Metropolis-within-Gibbs Markov chain is given as in the previous sections, then the transition operator T_θ is continuous with respect to θ in the following sense:

$$\int |T_\theta(x,dy) - T_{\hat{\theta}}(x,dy)| \leq C \cdot |\theta - \hat{\theta}|, \quad x \in X; \theta, \hat{\theta} > 0 \tag{6.45}$$

Furthermore $H(\theta,x)$ for our algorithm is continuous with respect to θ and satisfies the following conditions:

$$\begin{aligned} |H(\theta,x)| &\leq C, \quad x \in X & (6.46)\\ |T_\theta H(\theta,x) - T_{\hat{\theta}} H(\hat{\theta},x)| &\leq \bar{C} \cdot |\theta - \hat{\theta}|, \quad x \in X; \theta, \hat{\theta} > 0 & (6.47) \end{aligned}$$

for some constant C, $\bar{C}$.

Proof: Given a vector of underlying random numbers $\{z\}$ defined in some domain $B = \{B_1, B_2, ..., B_n\} \subset R^d$, the n-th transition kernel corresponding to the j-th coordinate for the Metropolis-within-Gibbs algorithm is given by:

$$\begin{aligned}
[T_\theta]_n^j &\equiv [T_\theta]_n^j(x_n; B_j \times \{x_n^1, x_n^2, ..., x_n^{j-1}, x_{n-1}^{j+1}, ..., x_{n-1}^d\}) \\
&= \int_{B_j} q_j(z^j - x_{n-1}^j) \cdot \alpha_n^j(z^j, \theta) \cdot dz^j \\
&\quad + I_{B_j} \int_R q_j(x_{n-1}^j - z^j) \cdot (1 - \alpha_n^j(z^j, \theta)) \cdot dz_j
\end{aligned} \tag{6.48}$$

where $\alpha_n^j(z^j, \theta)$ is given by:

$$\begin{aligned}
\alpha_n^j(z^j, \theta) &= 1 \wedge \frac{\pi_\theta(x_n^1, x_n^2, ..., x_n^{j-1}, z^j, x_{n-1}^{j+1}, ..., x_{n-1}^d)}{\pi_\theta(x_n^1, x_n^2, ..., x_n^{j-1}, x_{n-1}^j, x_{n-1}^{j+1}, ..., x_{n-1}^d)} \\
&= 1 \wedge C \cdot h(x, z^j) \cdot e^{-\theta \cdot r(x, z^j)}
\end{aligned} \tag{6.49}$$

where $r(x, z_j) = \max(V(x_n^1, x_n^2, ..., x_n^{j-1}, z^j, x_{n-1}^{j+1}, ..., x_{n-1}^d), v)$
$- \max(V(x_n^1, x_n^2, ..., x_n^{j-1}, x_{n-1}^j, x_{n-1}^{j+1}, ..., x_{n-1}^d), v)$, and $h(x, z^j)$ is the ratio of two gaussian densities.

In this algorithm, only one coordinate is updated at a time. Thus this is really an one dimensional formula with respect to z_j. In order to investigate the continuity of $[T_\theta]_n^j$ with respect to the parameter θ, let us take another $\hat{\theta}$, and compute the following:

$$\begin{aligned}
\int_{B_j} |[T_\theta]_n^j - [T_{\hat{\theta}}]_n^j| dx_{n-1}^j &\leq \int_{B_j} |\int_{B_j} q_j(z^j - x_{n-1}^j) \cdot A_n^j(\theta, \hat{\theta}) \cdot dz^j| dx_{n-1}^j \\
&\quad + \int_{B_j} |I_{B_j} \int_R q_j(z^j - x_{n-1}^j) \cdot B_n^j(\theta, \hat{\theta}) \cdot dz_j| dx_{n-1}^j
\end{aligned} \tag{6.50}$$

where

$$A_n^j(\theta, \hat{\theta}) = \alpha_n^j(z^j, \theta) - \alpha_n^j(z^j, \hat{\theta}) \tag{6.51}$$

$$B_n^j(\theta, \hat{\theta}) = 1 - \alpha_n^j(z^j, \theta) - (1 - \alpha_n^j(z^j, \hat{\theta})) = -A_n^j(\theta, \hat{\theta}) \tag{6.52}$$

If both θ, and $\hat{\theta}$ result in acceptance region (i.e. $\alpha_n^j(z^j, \theta) = 1$ and $\alpha_n^j(z^j, \hat{\theta}) = 1$), then $\int_{B_j} |[T_\theta]_n^j - [T_{\hat{\theta}}]_n^j| dx_{n-1}^j = 0$ (Constant). So, the continuity is obvious. Assume that both θ, and $\hat{\theta}$ result in rejection region. It is clear from the expression (6.49) that $\alpha_n^j(z^j, \theta)$ is continuous with respect to θ.

Moreover, since the function $e^{-\theta\, r(x, z_j)}$ is positive and bounded from above, we can find a constant $\hat{C}$, and $\bar{C}$ such that:

$$|A_n^j(\theta, \hat{\theta})| = |B_n^j(\theta, \hat{\theta})| = \hat{C}|e^{-\theta\, r(x, z_j)} - e^{-\hat{\theta}\, r(x, z_j)}| \leq \bar{C}|\theta - \hat{\theta}| \tag{6.53}$$

So, plugging these into the expression (6.50) gives:

$$\int_{B_j} |[T_\theta]_n^j - [T_{\hat{\theta}}]_n^j| dx_{n-1}^j \leq \int_{B_j} |\int_{B_j} q_j(z^j - x_{n-1}^j) \cdot \bar{C}|\theta - \hat{\theta}| \cdot dz^j| dx_{n-1}^j$$

$$
\begin{aligned}
&+\int_{B_j}|I_{B_j}\int_R q_j(z^j-x^j_{n-1})\cdot\bar{C}|\theta-\hat{\theta}|\cdot dz_j|dx^j_{n-1} \\
=\;& 2\bar{C}|\theta-\hat{\theta}| \qquad (6.54)
\end{aligned}
$$

Thus, the continuity follows. Since $H(x,\theta)=sign(V-v)$ (a constant function with respect to θ), and $|H(x,\theta)|\leq 1$, the continuity and the boundedness are obvious. The continuity of $[T_\theta]^j_n\cdot H(x,\theta)$ can be proved by the identical argument above since $|H(x,\theta)|$ is bounded by 1.

QED

Thus far we have proven that our chain satisfies the θ continuity theorem. We shall prove that the geometric ergodicity and the θ continuity impliy the Poisson condition for our chain. The proof is based on Delyon ([14]), but slighly altered for our MCMC chain.

Thorem 6.2.8 (Poisson Theorem)
Let us assume that the Random Scan Metropolis-within-Gibbs Markov chain with transition T_θ is geometrically ergodic, and satisfies the θ Continuity Theorem. Let us further assume that $\theta\in K$ for some compact set $K\subset R^+$. Then T_θ satisfies the Poisson condition.

Proof: Since T_θ is aperiodic and ergodic, T_θ is a contraction operator with 1 being the only eigenvalue (the isolated point of the spectrum) of T_θ on the unit circle (Sokal, p6, [64]). Therefore, the spectrum $\sigma(T_\theta)$ of T_θ can be decomposed into $\sigma(\pi_\theta)=1$ and the rest $\sigma(R_\theta)$ in such a way that a rectifiable, simple closed curve Γ can be drawn in the complex plane so that $\sigma(\pi_\theta)$ is in its interior, and $\sigma(R_\theta)$ is in its exterior (Kato, p.178, [49]). Furthermore, T_θ can be represented as: $T_\theta=\pi_\theta+R_\theta$. Furthermore, π_θ can be explicitly represented as:

$$
\pi_\theta=-\frac{1}{2i\pi}\int_\Gamma(T_\theta-zI)^{-1}zdz \qquad (6.55)
$$

where Γ is a contour near 1 in the complex plane, and R_θ is called the 'Resolvent' of the operator T_θ. R_θ is an operator with spectral radius ($\rho=lim\|R^n_\theta\|^{1/n}$) strictly less than 1. Because of the geometric ergodicity, the above representation for π_θ is finite. Since T_θ is continuous with respect to the total variation norm $\|\|$, the decomposition $T_\theta=\pi_\theta+R_\theta$ holds continuously with respect to θ (Kato, p.212, [49]). Thus for $\bar{\theta}$ close enough to θ (so that Γ is chosen independently of $\bar{\theta}$), we have:

$$
\begin{aligned}
\|\pi_{\bar{\theta}}-\pi_\theta\| &\leq \sup_{z\in\Gamma}\|(T_{\bar{\theta}}-zI)^{-1}-(T_\theta-zI)^{-1}\| \\
&\leq \sup_{z\in\Gamma}\|(T_{\bar{\theta}}-zI)^{-1}(T_{\bar{\theta}}-T_\theta)(T_\theta-zI)^{-1}\| \\
&\leq \sup_{z\in\Gamma}\|(T_{\bar{\theta}}-zI)^{-1}\|\|T_{\bar{\theta}}-T_\theta\|\sup_{z\in\Gamma}\|(T_\theta-zI)^{-1}\| \\
&\leq C\|T_{\bar{\theta}}-T_\theta\|\leq\bar{C}|\bar{\theta}-\theta|
\end{aligned}
$$

Let us define v_θ as:

$$v_\theta = \sum_{i=0}^{\infty}(T_\theta - \pi_\theta)^i H_\theta = \sum_{i=0}^{\infty}(T_\theta^i - \pi_\theta)H_\theta \tag{6.56}$$

By the geometric ergodicity and the boundedness of $H_\theta = sign(V - v)$, we have : $\|v_\theta\| \leq \sum_{i=0}^{\infty} \|(T_\theta^i - \pi_\theta)H_\theta\| \leq \sum_{i=0}^{\infty} C_\theta^{'}\rho^n = C_\theta < \infty$, $(\rho < 1)$. Since θ is in the compact set K, and the spectrum of T_θ is upper semicontinuous, the above constant C_θ can be made independent of θ. Therefore, we have $\|v_\theta\| \leq C < \infty$, so (2) of the Poisson Condition is satisfied. Furthermore, since $\pi_\theta T_\theta = \pi_\theta$, we have the relation: $(I - T_\theta)v_\theta = (I - T_\theta + \pi_\theta - \pi_\theta)\sum_{i=0}^{\infty}(T_\theta^i - \pi_\theta)H_\theta = \sum_{i=0}^{\infty}(T_\theta^i - \pi_\theta)H_\theta - \sum_{i=1}^{\infty}(T_\theta^i - \pi_\theta)H_\theta - \pi_\theta\sum_{i=0}^{\infty}(T_\theta^i - \pi_\theta)H_\theta = (I - \pi_\theta)H_\theta - \pi_\theta\sum_{i=0}^{\infty}(T_\theta - \pi_\theta)^i H_\theta = H_\theta - \pi_\theta H_\theta$ which is the Poisson equation. Therefore (1) of the Poisson condition is satisfied. Since $\pi_\theta T_\theta = \pi_\theta$, we have $\pi_\theta v_\theta = \pi_\theta \sum_{i=0}^{\infty}(T_\theta^i - \pi_\theta)H_\theta = 0$. By using these relations, we have the following equalities:

$$\begin{aligned}
(I - R_\theta)(v_{\bar\theta} - v_\theta) &= (I - T_\theta + \pi_\theta)(v_{\bar\theta} - v_\theta) \\
&= -(I - T_\theta)v_\theta + (I - T_\theta)v_{\bar\theta} + \pi_\theta v_{\bar\theta} - \pi_\theta v_\theta \\
&= -(I - T_\theta)v_\theta + (I - T_{\bar\theta} + T_{\bar\theta} - T_\theta)v_{\bar\theta} + \pi_\theta v_{\bar\theta} \\
&= -(I - T_\theta)v_\theta + (I - T_{\bar\theta})v_{\bar\theta} + (T_{\bar\theta} - T_\theta)v_{\bar\theta} + \pi_\theta v_{\bar\theta} \\
&= -(I - T_\theta)v_\theta + (I - T_{\bar\theta})v_{\bar\theta} + (T_{\bar\theta} - T_\theta)v_{\bar\theta} + \pi_\theta v_{\bar\theta} - \pi_{\bar\theta} v_{\bar\theta} \\
&= -(H_\theta - \pi_\theta H_\theta) + (H_{\bar\theta} - \pi_{\bar\theta} H_{\bar\theta}) + (T_{\bar\theta} - T_\theta)v_{\bar\theta} + (\pi_\theta - \pi_{\bar\theta})v_{\bar\theta}
\end{aligned}$$

Since the spectral radius for the residual operator R_θ is strictly less than 1, we have $\|(I - R_\theta)^{-1}\| < \infty$. Thus,

$$\begin{aligned}
\|v_{\bar\theta} - v_\theta\| &\leq C\|(H_{\bar\theta} - \pi_{\bar\theta} H_{\bar\theta}) - (H_\theta - \pi_\theta H_\theta) + (T_{\bar\theta} - T_\theta)v_{\bar\theta} + (\pi_\theta - \pi_{\bar\theta})v_{\bar\theta}\| \\
&\leq \bar{C}|\bar\theta - \theta|
\end{aligned}$$

since H_θ, $\pi_\theta H_\theta$, T_θ, π_θ are all continuous with respect to θ. Therefore, we finally have:

$$\|T_{\bar\theta} v_{\bar\theta} - T_\theta v_\theta\| \leq \|T_{\bar\theta}(v_{\bar\theta} - v_\theta)\| + \|(T_{\bar\theta} - T_\theta)v_\theta\| \leq C|\bar\theta - \theta|$$

by the continuity of T_θ and v_θ under the norm $\|\|$, and the boundedness of $T_{\bar\theta}$ and v_θ. This satisfies (3) of the Poisson Condition.

QED

Therefore our chain satisfies the Poisson condition. In the next section, we shall state a general Robbins-Monro Theorem.

6.2.4 Robbins-Monro Theorem for MCMC

Thus far, we have shown that our chain (Random Scan Metropolis-within-Gibbs) satisfies the Luapunov condition, the geometric ergodicity, the Poisson condition, and the θ-continuity theorem. Since our parameter family θ is contained in some compact set $K \subset R^+$, this is enough to prove almost sure convergence for our adaptive Robbins-Monro algorithm. We shall provide a general Robbins-Monro theorem for the MCMC case. The proof is based again on Delyon ([14]). In his proof, arguments are made based on the projection (Chen's algorithm). However, for our MCMC case, since θ is contained in the compact set K, we shall provide a simpler version of the theorem for our MCMC chain.

Thorem 6.2.9 (Robbins-Monro Theorem: Markov Chain Case)
Let us assume that the algorithm (6.1) satisfies the following conditions:

(1) The function h of the algorithm satisfies the Lyapunov condition.

(2) γ_n satisfies the Slowly decreasing condition.

(3) θ is contained in a compact set K.

(4) The Markov Chain satisfies the Poisson Condition.

Then $\theta_n \to \theta^*$ such that $h(\theta^*) = 0$ almost surely.

Proof: All we have to prove is that the chain satisfies the A-stability condition. (1) of the A-stability condition is trivially satisfied by (3) of the hypothesis above. Next, since the chain satisfies the Poisson condition ($(I - T_\theta)v_\theta(x) = H(\theta, x) - h(\theta)$), we have:

$$\begin{aligned} H(\theta_n, X_{n+1}) &= h(\theta_n) + v_{\theta_n}(X_{n+1}) - T_{\theta_n} v_{\theta_n}(X_{n+1}) \\ &= h(\theta_n) + v_{\theta_n}(X_{n+1}) - T_{\theta_n} v_{\theta_n}(X_n) \;\; [\equiv> h(\theta_n) + \xi_{n+1}] \\ &+ T_{\theta_n} v_{\theta_n}(X_n) - T_{\theta_{n+1}} v_{\theta_{n+1}}(X_{n+1}) \;\; [\equiv> \bar{e}_n - \bar{e}_{n+1}] \\ &+ T_{\theta_{n+1}} v_{\theta_{n+1}}(X_{n+1}) - T_{\theta_n} v_{\theta_n}(X_{n+1}) \;\; [\equiv> e_{n+1}] \end{aligned}$$

ξ_{n+1} is a martingale increment part. Furthermore, $|T_\theta v_\theta(x)| \leq |v_\theta(x)| + |H(\theta, x)| + |h(\theta)| \leq C$ by (2) of the Poisson condition. Thus $\sum E[|\gamma_{n+1}\xi_{n+1}|^2 | F_n] \leq C \sum \gamma_{n+1}^2 < \infty$. Therefore, by the Martingale convergence theorem, $\lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i \xi_i$ exists almost surely. Thus (2) of the A-stability condition is satisfied. Next, since $|\bar{e}_n| \equiv |T_{\theta_n} v_{\theta_n}(X_n)| < \infty$ and θ_n is contained in the compact set K. The limit $\bar{e}_n$ converges to a point in the compact set K. Therefore $\lim_{n\to\infty} |\bar{e}_n - \bar{e}_{n+1}| = 0$. Furthermore, by (3) of the Poisson condition, and by the definition of the Robbins-Monro algorithm $\theta_{n+1} = \theta_n + \gamma_n H(\theta_n, X_n)$, we have:

$$\begin{aligned} |e_{n+1}| &= |T_{\theta_{n+1}} v_{\theta_{n+1}}(X_{n+1}) - T_{\theta_n} v_{\theta_n}(X_{n+1})| \\ &\leq \bar{C}|\theta_{n+1} - \theta_n| = \bar{C}\gamma_n |H(\theta_n, X_n)| \leq \hat{C}\gamma_n \end{aligned}$$

Thus $\lim_{n\to\infty} |e_{n+1}| = 0$. So (3) of the A-stability condition is satisfied. Therefore, the A-stability condition is satisfied almost surely. Thus, by the A-stability condition together with (1) and (2) of the hypothesis of this theorem, the assertion of the Theorem 6.1.6 (General Convergence Theorem: ODE Part) holds true almost surely. That is $\theta_n \to \theta^*$ such that $h(\theta^*) = 0$ almost surely.

QED

The above proof is essentially the same for the case when Chen's algorithm is applied. Thus for our MCMC chain, the assertion still holds when Chen's algorithm is implemented.

6.3 Large Deviation Analysis for MCMC: Asymptotic Efficiency

In chapter 3, we have shown the asymptotic efficiency for both gaussian twist and exponential twist (i.i.d. case). The similar argument can be made for the MCMC cases as well if the chain is geometrically ergodic. We shall show a heuristic argument based on the autocorrelation time, and the . The estimator $\hat{p} = \frac{\hat{x}}{\hat{y}} = \frac{\frac{1}{N} I_{V_i \leq v} L_i}{\frac{1}{N} L_i}$ for the VAR probability involves estimators for both the numerator $\hat{x}$ and the denominator $\hat{y}$ (normalization constant) where L_i is the likelihood ratio. The expression for the variance $\sigma(\hat{x})^2$ is given by: $\sigma(\hat{x})^2 = \frac{\sigma(\hat{x}_{(i.i.d)})^2}{\tau_x}$. Similarly, we have: $\sigma(\hat{Y})^2 = \frac{\sigma(\hat{Y}_{(i.i.d)})^2}{\tau_y}$. $\sigma(\hat{X}_{(i.i.d)})^2$ and $\sigma(\hat{Y}_{(i.i.d)})^2$ are the variance for the numerator and the denominator respectively assuming that the sampling is done independently. τ_x, and τ_y are the autocorrelation time for the numerator and the denominator respectively. If the fluctuation of the denominator is small compared with the denominator (this is heuristic, but can be done algorithmically, section 5.5.3). Then we can make the first order estimate for the variance for $\hat{p}$ as $\sigma(\hat{p})^2 = \sigma(\hat{p}_{(i.i.d)})^2 \tau_x$. Let us assume that the VAR region B_v is a convex region with small curvature at the local minima β. By the results from the appendix C, the variance for $\sigma(\hat{p}_{(i.i.d)})^2$ decreases like $\sim \frac{e^{-|\beta|^2}}{|\beta|}$ as $|\beta| \to \infty$. Therefore, as the quantile becomes more extreme (meaning $\beta \to \infty$), $\sigma(\hat{p})^2$ decreases like $\sim \frac{\tau_x e^{-|\beta|^2}}{|\beta|}$ as $|\beta| \to \infty$. For any MCMC, $\tau > 1$. However, if the MCMC chain is geometrically ergodic, all the covariance terms are bounded by $\rho < 1$. Therefore, the dependency decreases geometrically. Thus, both τ_x and τ_y are expected to be reasonably small and finite ($< C$) (τ_x may increase with dimension). Thus for a fixed dimension, as $\beta \to \infty$, $\sigma(\hat{p})^2$ still drops like $\sim \frac{e^{-|\beta|^2}}{|\beta|}$. Therefore, our MCMC chain is asymptotically efficient. The same argument can be made for the multiple local minima cases as in chapter 3 since τ_x is bounded.

Chapter 7

Application to Computing VAR for Path-Dependent Options, and Pricing Basket Options

7.1 Introduction

Path dependent option is defined, as the name suggests, any contingent claim whose payoff depends on the past path taken by the underlying security(ies). The typical examples are Asian option, Lookback option, and most commonly American option. There are many papers suggesting pricing these securities. However, none talk about computing VAR for portfolio containing those securities. In the next sections, we shall take up these three path-dependent cases for computing VAR. Furthermore, we will show how our method can be used to effectively price basket options for highly out-of-the-money basket options. This can be used more like as an insurance purpose.

7.1.1 Asian Option

Asian option has been actively traded especially in fereign exchange market. It is faous because one can hedge relatively cheaper than headging each foreign exchange products. Asian option generally fall into two categories - Geometric Asian option, and Arithmetic Asian option. Geometric Asian option is a option on the product average of the past d prices for a single underlier given by $(\prod_{i=1}^{d} S(t_i))^{1/d}$. Similarly, an Arithmetic Asian option is an option on the arithmetic average of the past d prices for a single underlier given by $(\sum_{i=1}^{d} \frac{S(t_i)}{d})$. The former is known to admit a Black-Scholes type analytic solution, so we shall focus on the latter - the Arithmetic Asian option. The general formulae are as follows:

$$Asian\_Call(T) = E[e^{-rT}(\sum_{i=1}^{d} \frac{S(t_i)}{d} - K)^{+}] \tag{7.1}$$

$$Asian\_Put(T) = E[e^{-rT}(K - \sum_{i=1}^{d} \frac{S(t_i)}{d})^{+}] \tag{7.2}$$

Note that Asian option pricing translate its problem into m-dimensional problem because of the expression $\sum_{i=1}^{m} \frac{S(t_i)}{m}$. This is a discrete state path-dependent option. For pricing, there are two approaches one can take. One is based on correlation, the other is based on independence increment (Martingale) approach. Regarding the correlation method, it is well known that the correlation between any two given Brownian motions B_s, B_t is computed as:

$$Corr(B_s, B_t) = \frac{Cov(B_s, B_t)}{Var(B_s) \cdot Var(B_t)} = \frac{\sqrt{s \wedge t}}{s \cdot t} = \sqrt{\frac{s}{t}} \tag{7.3}$$

if $s \leq t$.
For the Asian option, each underlier risk factor $S(t_i)$ follows d-dimensional lognormal random process which can be the translated into multivariate correlated normal process x_i by simple change of variable componentwise as : $S(t_i) = S_i(0)e^{\sigma_i \sqrt{t_i} x_i + (r - \frac{\sigma_i^2}{2})t_i}$. Each x_i can be thought as the path taken by the Brownian underlier at time t_i. Therefore, the risk factor vector $(X_1, X_2, ...X_d) = (x_1\sqrt{t_1}, x_2\sqrt{t_2}, ...x_d\sqrt{t_d})$ has a positive definite correlation matrix Σ where $\sigma_{i,j} = \sqrt{\frac{t_i \wedge t_j}{t_i \vee t_j}} \leq 1$ where each x_i is a standard normal variable.

Thus far, we have explained the dynamics behind Asian option. We shall explain how one can compute VAR for Asian option. Let us take an Asian call option. If we are to use plain vanilla Monte Carlo, we just generate i.i.d. correlated standard normal variables x with correlation matrix Σ. Then construct $S(t_i)$, and compute $\sum_{i=1}^{d} \frac{S(t_i)}{d}$. Monte Carlo calls for repeating these steps thousands of times. When it is done, the average of the payoff $V = e^{-rT}(\sum_{i=1}^{d} \frac{S(t_i)}{d} - K)^{+}$ will give the premium or price. The VAR is a quantile of the distribution of the quantity V. Our algorithm can be used for this frame work. V is given as above. If we know what the p-th quantile VAR v for this security from previous day, and assuming that the position did not change overnight, then we can just take the target density as: $fe^{-\theta max(V,v)}$. Therefore, the algorithm in the chapter 5 can be used. Its simulation error can be computed by the Method Batched Means from chapter 5 as well.

In the previous paragraphs, we have explained correlation approach. There is another formulation which is more efficient than this method. It is based on the Martingale increment approach. Brownian motion has the characteristic that the increment is independent of the previous state. That is B_{t-s} is independent of B_s. Therefore unlike the approach we took in the previous paragraph, we can construct independent random variables $\hat{x}$ as:

$$\bar{x} = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_d\} = \{\hat{x}_1\sqrt{t_1}, \hat{x}_2\sqrt{t_2 - t_1}, ...\hat{x}_d\sqrt{t_d - t_{d-1}}\} \tag{7.4}$$

Once this is done, then each S_{t_i} cab be computed by the following product: $S(t_i) = S(0)e^{(r-\frac{\sigma_i^2}{2})t_i} \prod_{k=1}^{i} e^{(\sigma_i \bar{x}_k)}$. Notice that each $S(t_i)$ can be computed recursively from the previous $S(t_{i-1})$ since:

$$S(t_i) = S(0)e^{(r-\frac{\sigma_i^2}{2})t_i} \prod_{k=1}^{i-1} e^{\sigma_i \bar{x}_k} \cdot e^{\sigma_i \bar{x}_i} = S(t_{i-1})e^{\sigma_i \bar{x}_i} \tag{7.5}$$

This is really equivalent to the correlated version. However, this formulation has a computational advantage. The creation of next random variable $S(t_i)$ is multiplicative (or additive in terms of normal deviate). Therefore, the operation cost is only d (linear with dimension) for this generation. Whereas, for the correlation case, the operation cost is $\frac{d(d+1)}{2}$ (quadratic). This is because of the cross multiplication with correlation matrix for each run which could cost computationally if the dimension huge.

For this formulation, we can apply the algorithm to $\bar{x}$ with $fe^{-\theta max(V,v)}$. Since this is not correlated, componentwise update of Metropolis-within-Gibbs algorithm is easier to apply as well. Experimental results will be shown in the next chapter. In the next section, Lookback option will be taken up. For this option, we can use the similar argument as explained in this section.

7.1.2 Lookback Option

Lookback option is also another example of path-dependent option. This option pays premium based on the maximum or the minimum of the values attained by the underlying securities for the life of the option. Its price is given in general as follows:

$$Lookback\_Call(T) = E[e^{-rT}(\max_{t\in[0,T]}(S(t)) - K)^+] \tag{7.6}$$

The formula for the put is similar except $(K - \max_{t\in[0,T]}(S(t)))^+$ inside the expectation. Unlike the Asian option where time grids are already given, the Lookback option looks for the entire past path continuously. In practice, it is impossible to do Monte Carlo on the infinite dimension. Therefore we need to make some assumption about the number of time grids for this option. This is called discretization. It is a standard technique in numerical PDE methods. We shall assume that there are d (large enough) time grids (or dimension) so that the option price for the Lookback can be approximated by:

$$Lookback\_Call(T) = [e^{-rT}(\max_{i=1,2,\ldots d}(S(t_i)) - K)^+] \tag{7.7}$$

Once this formulation is done, then the underlying risk factors $S(t_i)$ can be constructed either using the correlation method or the Martingale increment method as described in the previous section. We take V to be $V = e^{-rT}(\max_{i=1,2,\ldots d}(S(t_i)) - K)^+$, and construct our usual target density $fe^{-\theta max(V,v)}$. Then our Metropolis-within-Gibbs algorithm can be used to compute VAR. We had made heuristic argument for making grid for our case. For Lookback option, we believe this approach is appropriate. However, there is another path-dependent option where making grid has to be chosen carefully. This is pricing or computing VAR for American option. We shall take up this isse next.

7.1.3 American Option

American option is a widely traded option in the securities and exchange market. It is a contract that one has the right to buy (Call) or sell (Put) an underlying security any time during the life of the option. It is path-dependent because the entire past history of the underling security has be investigated. Furthermore, this does not admit analytical solution as it is a free-boundary problem, and it has traditionally been computed by various PDE methods (either binomial tree, trinomial tree, or Crank-Nicholson). Recently, many papers have emerged (Barraquad [53], Glasserman [33], etc) for pricing such product by Monte Carlo. There have not been any papers published on computing VAR for American option. Most practitioners treat American as European, or some analytic approximation (Whaley) to compensate for this situation. On the other hand, performing numerical PDE thousands of times just for computing VAR is not practical either. In this section, we will formally show how one can compute VAR for American option. The word of caution is needed here. Our method in the next section is not exact. It is an approximation by Monte Carlo. Therefore, it is subject to some errors just like any numerical methods

American option is described as:

$$American\_Call(T) = \sup_{\tau \in H} E[e^{-r\tau}(S(\tau) - K)^+] \tag{7.8}$$

where τ is a stopping time, and H is a set of all possible optimal exercise stopping times. The expression for put is similar except the expression inside the expectation is $e^{-r\tau}(K - S(\tau))^+$. It is well known that American Call with no dividend has no early exercise boundary, therefore it is equal to European price. We shall exclude these trivial cases. Just like any other path-dependent options in the previous section, we need to make some grid assumptions. Thus, what we propose is the following algorithm. At first, we run numerical PDE (either binomial or trinomial tree) just once. When the tree is finished, we have a set of early exercise periods $(t_1, t_2, ..., t_{d-1})$ and associated early exercise boundaries $(X_1^1, X_1^2, ...X_2^1, X_2^2, ...X_{d-1}^1...)$. Next, we take the minimum of the early exercise boundaries corresponding to each of the early exercise periods, say $(\bar{X}_1, \bar{X}_2, ..., \bar{X}_{d-1})$. Therefore $(t_1, t_2, ...t_{d-1}, T)$ would be grid we take for our Monte Carlo simulation. Note that the numbers of gridding will be different depending on how fine the binomial or trinomial tree is performed. Also notice that this PDE method is run only once as a guidance for Monte Carlo simulation. Let us explain the algorithm in more detail;

Algorithm 7.1.1 ($M(RT^2)$ Algorithm for Computing VAR for American Option)

(1) Run Numerical PDE method only once (binomial or trinomial) to get a set of early exercise period $(t_1, t_2, ..., t_{d-1}, T)$ and corresponding early exercise boundaries $(\bar{X}_1, \bar{X}_2, ..., \bar{X}_{d-1}, 0)$

(2) Construct a set of risk factors $S(t_i)$ as in the previous section.

(3) For each exercise time grid t_i, compare $S(t_i)$ against X_i. If $S(t_i) \geq X_i$, then the early exercise boundary has been reached. Thus take $e^{-rt_i}(S(t_i) - K)^+$ as the American payoff value V, and go to the step (6).

(4) If $S(t_i) < X_i$, then the early exercise boundary has not been reached, so repeat the step (3) for the next grid.

(5) If the exercise boundary is not reached by the last time T, then take $e^{-rT}(S(T) - K)^+$ as the final American payoff value V and go to the step (6).

(6) Take this V, and initial θ_0, and start the Robbins-Monro algorithm described in chapter 5 with our familiar target density $\pi_\theta = fe^{-\theta max(V,v)}$.

(7) Repeat the above steps (2)-(6) for the rest of the simulation.

Notice that the if there is no early exercise, then $V = e^{-rT}(S(T) - K)^+$ which is a one dimensional problem with time $= T$. Thus this algorithm will be reduced to the European problem.

Thus taking this as a hint, anything that can be computed by the PDE method (as the worst case) can be applied here for our Monte Carlo algorithm. For example, products like mortgage backed securities (MBS), and any other fixed income securities with certain pre-payment can be analyzed similarly to compute its risk using our method. In this paper, we have focused on rare event simulation application to VAR. However, this has an application to pricing as well. This is for pricing basket option for far out-of-the-money option for insurance purposes. We shall take up this issue in the next section.

7.1.4 Pricing Basket Option on a Portfolio

So far we have focused on computing VAR. This is mainly because of the fact that our algorithm works well for rare event simulations, and not for pricing which requires sampling at the center of distribution. However, our algorithm can be applied to price far-out-of-the-money basket options. Typically, these are often sold for insurance purposes (to hedge against some extreme risks) such as for some structured portfolio or book of foreign exchange products. One may wants to price a put option where you have the right to sell the portfolio V if the portfolio value goes below $v = K$. If the this v is in the tail event, then we can apply the same algorithm in chapter 5 with the target density $\pi_\theta = fe^{-\theta max(V,v)}$. The pricing can be computed by the following expression with appropriate importance weights:

$$Basket\_Option\_Put\_Price(T) \approx \frac{\sum_{i=1}^{N}(v-V)^{+}e^{\theta max(V,v)}}{\sum_{i=1}^{N}e^{\theta max(V,v)}} \tag{7.9}$$

Its error can be computed by the Method of Batched Means as well. One can see the effectiveness of this algorithm by looking at the target density π_θ.

$$\pi_\theta = C \cdot fe^{-\theta max(V,v)} = \hat{C} \cdot fe^{-\theta max(V-v,0)} \tag{7.10}$$

The limiting distribution tends to be centered around the set $V = v$. The optimization is easier, and its concept makes sense since the approach is to treat V as a random number, not its underlying risk factors. The risk factors are the part of the system. But they are all put in the black box approach as this algorithm implicitly find the optimal point based on the behavior of the random number V.

7.2 Problem Formulation for VAR

We shall explain how we can use our MCMC algorithm in practice. For computing VAR probability for a portfolio V containing path-dependent options described in previous sections, one can basically translate each path-dependent securities into multi-dimensional European type securities with more dimensions. This will increase the dimension of the problem V. Suppose V contains d securities, and each security contains $D(d) > 1$ dimensions. Then the total effective dimension for simulating V would not be d, but $\sum_{i=1}^{d} D(i) > d$. One could use plain Vanilla i.i.d. Monte Carlo, or try to use i.i.d. based importance sampling described in chapter 3. However, as the dimension gets larger, setting up such optimization may not be practical. In addition, if the portfolio contains many local minima, then it will be harder. Thus, this is where the Metropolis Metropolis-within-Gibbs algorithm can play a role. Most portfolio managers know what the typical VAR number is either from experience (yesterday) or from existing trading portfolio similar to this. Let us call it v. Then one can follow the algorithm in chapter 5, without doing any extra calculations other than computing V. The density $f \cdot e^{-\theta \cdot max(V,v)}$ aids us in sampling most of the time near the VAR region. For path-dependent cases, one has to be careful that the dimension of the number of securities d, and actual risk factors $\sum_{i=1}^{d} D(i)$ are different. Therefore, in performing the Metropolis-within-Gibbs method, one has to make sure that the algorithm sweeps through all dimensions of the risk factors.

This sampling method is not faster than any i.i.d. based importance sampling (if one can do it) because of the autocorrelation time (usually from an order of 5 to 15) depending on dimension. Nevertheless, it is most likely better than the plain vanilla Monte Carlo method if the quantile is in the far tail of the distribution of portfolio value V and if the simulation identifies all possible local minima. This is demonstrated both experimentally as in the next chapter, and theoretically by the previous chapter (geometric ergodicity of our chain and the asymptotically efficiency of our method for the rare event simulation). Another important fact about this algorithm is that we can sample from any shape of distribution as long as the ergodicity and the θ continuity are satisfied (chapter 6). This algorithm allows extra degree of freedom. For example, some practitioners may prefer to sample from multivariate t-distribution, or other kind as f instead of the usual multivariate standard normals. Even for these cases, our algorithm can be applied by plugging these f into the density $\pi_\theta = f \cdot e^{-\theta \cdot max(V,v)}$.

On the other hand, for some extreme portfolios with multiple local minima, simulation may stuck in one local minimum. If this happens, our algorithm can fail due to missing some important local minima. Moreover, the autocorrelation time can grow significantly as the dimension becomes larger for some portfolios. Then, the variance ratio can go below 1 (less effective than the regular Monte Carlo method if it is possible to do). Therefore, one has to be aware of the potential shortcoming of this method. Nevertheless, if i.i.d. based sampling is impossible, then this algorithm is the best candidate for sampling rare event scenarios. In the next chapter, several experimental results will be shown.

Chapter 8

Numerical Results (Markov Chain Case)

We have taken several examples ranging from simple portfolio to portfolio containing path dependent options. In this section, we shall desecribe series of experimental results which demonstrate that our Metropolis-within-Gibbs algorithm is working as it should. We tested both Random Scan Metropolis-within-Gibbs and Sequential Scan Metropolis-within-Gibbs algorithms. Their results were almost identical in that the differences are within 10 %. For this reason, the resulting tables below were shown for the Sequential Scan Metropolis-within-Gibbs algorithms only. Let us first take up simpler example. We have tested portfolio containing calls and puts for various combinations. Then we constructed examples where there are multiple local minima. Next, we took up examples of path dependent options - Asian, Lookbacks, and American options. For all cases, our method worked well.

The following tables show variance ratios and autocorrelation time for various test portfolios. All VAR probabilities in this tests are computed around $p = 0.01$ (10 day 99 % VAR). Unless otherwise stated, the parameters for all experiments for this section are set as follows: Risk Free Interest Rate = 0.05, Dividend = 0 (Except for American option where it is set to 0.043), Time to maturity = 0.05, Strike = 90, Volatility = 0.3 (Except for path dependent cases, we set to 0.21) Number of Iterations = 500,000 (Batch Size = 1000, Number of batches = 500) for dimension less than 10. Number of Iterations = 1,000,000 (Batch Size = 1000, Number of batches = 1000) for dimension higher than 10. The first set of portfolio is described as follows:

(Portfolio Description)

1 10 short cash positions in 2 dimensions correlated with all offdiagonal element = 0.2.

2 10 short call positions in 4 dimensions correlated with all offdiagonal element = 0.2.

3 10 short call positions in 8 dimensions correlated with all offdiagonal element = 0.4.

4 10 short call delta hedged with put positions in 4 dimensions correlated with all offdiagonal element = 0.2.

5 10 short call positions in 4 dimensions correlated with all offdiagonal element = 0.2.

6 10 long call, 5 long put, and 1 cash positions in 6 dimensions correlated with all offdiagonal element = 0.5 .

7 Butterfly positions in 2 dimensions uncorrelated: one long call with strike = 90, one long call with strike = 110, and 2 short call with strike =100.

8 The same butterfly as (5) with correlation with all offdiagonal = 0.4.

[**Table 8.1: MCMC Metropolis-within-Gibbs Algorithm Variance Ratio Test Results**]

Portfolio	Probability	Variance Ratio	Autocorrelation
1	9.234562E-003	4.567343E+000	8.567894E+000
2	9.678242E-003	4.012319E+000	8.233451E+000
3	1.012344E-002	3.871734E+000	1.224605E+001
4*	1.099434E-002	3.854445E+000	8.342344E+000
5	1.045434E-002	3.9445675E+000	8.5675997E+000
6	1.053231E-002	4.0454344E+000	8.955656E+000
7*	1.012344E-002	3.455656E+000	8.024348E+000
8*	9.342324E-003	4.0234557E+000	8.735677E+000

(* is a portfolio with multiple local minima)

Notice that the variance ratio is around 4 which is significantly lower than i.i.d. cases where the order was around 40-200. This is due to the autocorrelation time associated with MCMC where the autocorrelation times are around 10.

Next we have tested various path-dependent options including Asian, Lookbacks, and American options. Here we have implemented the algorithm in chapter 5 for these path-dependent options. The comparison with any of the i.i.d. based importance sampling was not done because of the difficulty of applying any of these methods into these portfolios. This experiments demonstrate that even with extremely difficult cases, MCMC sampling seems to perform well.

Unless otherwise stated, all parameters (volatility, rate, etc) are the same as before. The detail descriptions of these portfolios are set as follows:

(Portfolio Description)

1 10 long Asian call with strike = 100, the time grid size equal to 10 equal intervals with the final maturity time = 0.05. This results in 10 effective dimensions.

2 10 short Asian put with strike = 100. The effective dimension is the same as the above.

3 10 long Lookback call positions with effective dimension of 10. The gridding was done the same way as the Asian case above.

4 10 long Lookback call positions with effective dimension of 30. The gridding was done the same way as the Asian case above.

5 10 long American call option with effective dimension of 5. The exercise times reported by the binomial tree are (0.03, 0.035, 0.04, 0.045, 0.05). The corresponding early exercise boundaries are (109.269, 107.650, 109.252, 107.634, 0).

6 10 long American call option with effective dimension of 17. The gridding for the binomial tree was increased so that it resulted in 17 exercise times. They are = (0.018, 0.020, 0.022, 0.024, 0.026, 0.028, 0.030, 0.032, 0.034, 0.036, 0.038, 0.04, 0.042, 0.044, 0.046, 0.048, 0.050). The corresponding exercise boundaries are = (108.790, 107.770, 108.784, 107.764, 108.778, 107.757, 108.771, 107.751, 108.764, 107.744, 108.758, 107.738, 106.727, 107.731, 106.721, 107.725, 0).

7 Basket option containing the American option in (6) plus 10 short European calls in 2 dimensions, 5 long European puts in the same dimensions as the calls, 5 digital call in another 2 dimensions , and 6 digital puts in the same dimensions as the digital calls. The blockwide correlation is all set to 0.5. This results in the effective dimension of 17 + 2 + 2 = 21.

8 Basket option containing the American option in (6) plus separate Butterfly positions in 6 dimensions block correlated with correlation = 0.5: one long call with strike = 90, one long call with strike = 110, and 2 short calls with strike =100.

[**Table 8.2 MCMC Metropolis-within-Gibbs Algorithm Variance Ratio Test Results**]

Portfolio	Probability	Variance Ratio	Autocorrelation
1	9.798877E-003	3.427232E+000	1.544545E+001
2	9.136089E-003	3.012343E+000	1.685677E+001
3	1.142574E-002	3.252323E+000	1.532123E+001
4	9.909934E-003	2.042234E+000	3.114677E+001
5	8.997843E-003	4.566678E+000	8.123445E+000
6	1.035561E-002	3.023355E+000	1.912424E+001
7	8.012344E-003	2.484545E+000	2.342456E+001
8*	7.342324E-003	2.364311E+000	2.645677E+001

(* is a portfolio with multiple local minima)

We can see the increases in autocorrelation time (over 10), and thus smaller variance ratios for these cases. Yet, even for these portfolios with many dimensions, the Metropolis Algorithm still seems to outperform the regular Monte Carlo method. It is sometimes instructive to show the dynamics of our Componentwide Metropolis algorithm in several pictures. In the following pictures, evidences are shown that our target density π_θ is correctly sampling all local minima for butterfly. For illustrative purposes, we have tested various 2 dim butterfly examples for graphic presentation purposes.

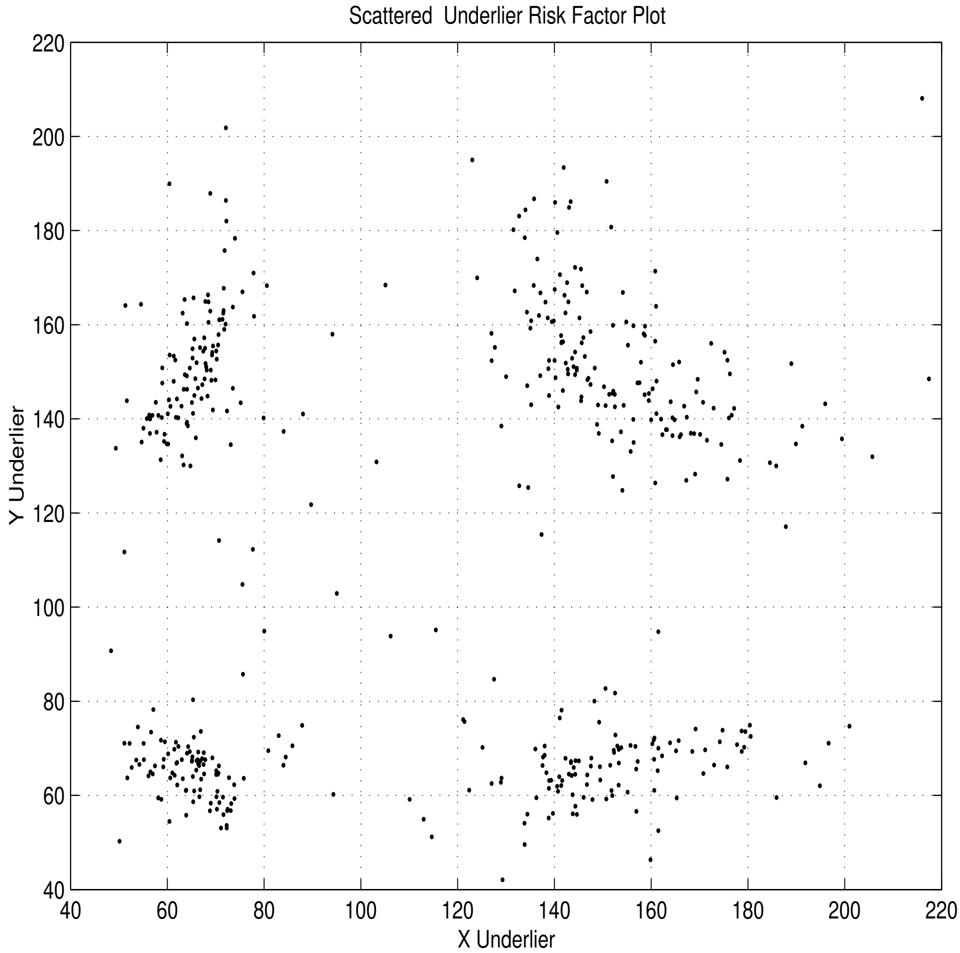

Figure 8.1: (Scattered Plot for the Generated Underliers: 2 Dimensional Butterflies Uncorrelated) Notice that this algorithm correctly samples all 4 local minima as we can see that there are four groups of humps each corresponding to local minima. We have tested up to 6 dimensions, our algorithm still seems to work even though in our case there are $2^6 = 64$ local minima.

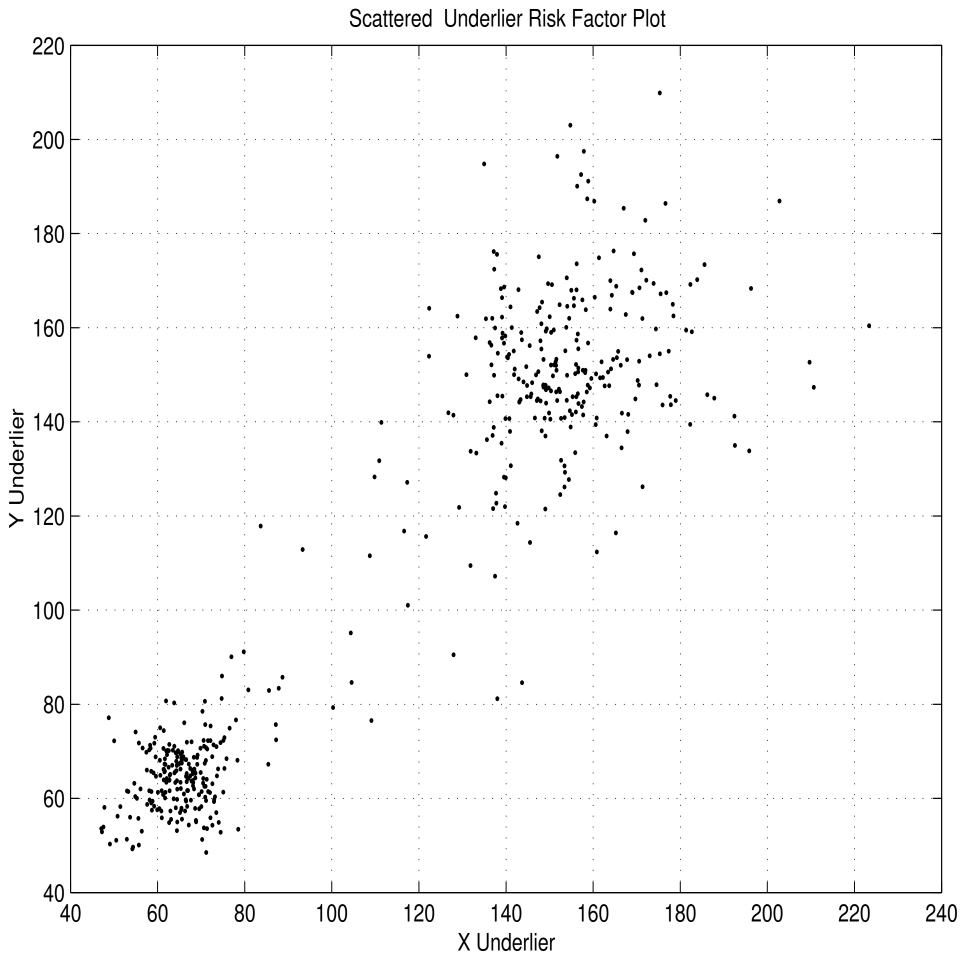

Figure 8.2: (Scattered Plot for the Generated Underliers: 2 Dimensional Butterflies Correlated with the Diagonal = 0.5) We can notice that the strong correlation has the effect of collapsing number of local minima into fewer local minima. This point makes sense based on principal component axis analysis. This algorithm still does a good job in sampling local minima for this case.

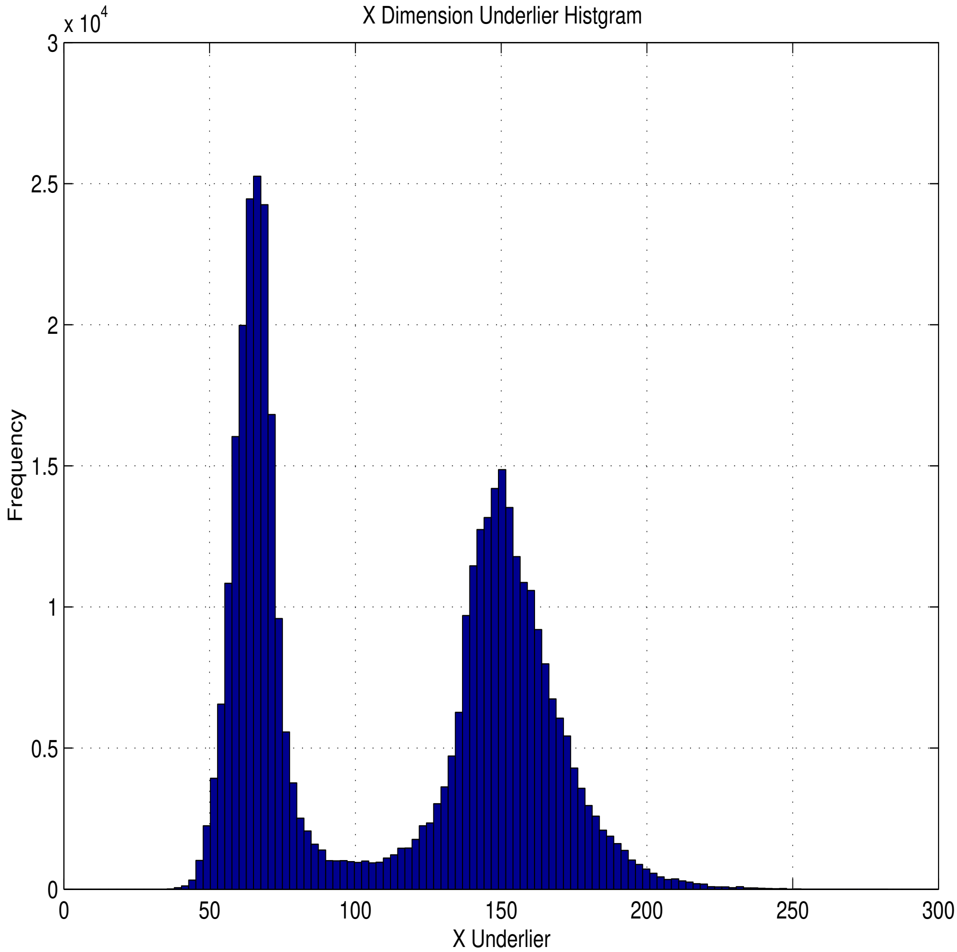

Figure 8.3: (The Histgram for the X Underlier: Butterfly Correlated with the Diagonal = 0.5) This corresponds to the same portfolio taken up in the previous page. The picture clearly indicates that π_θ is peaked near the VAR region.

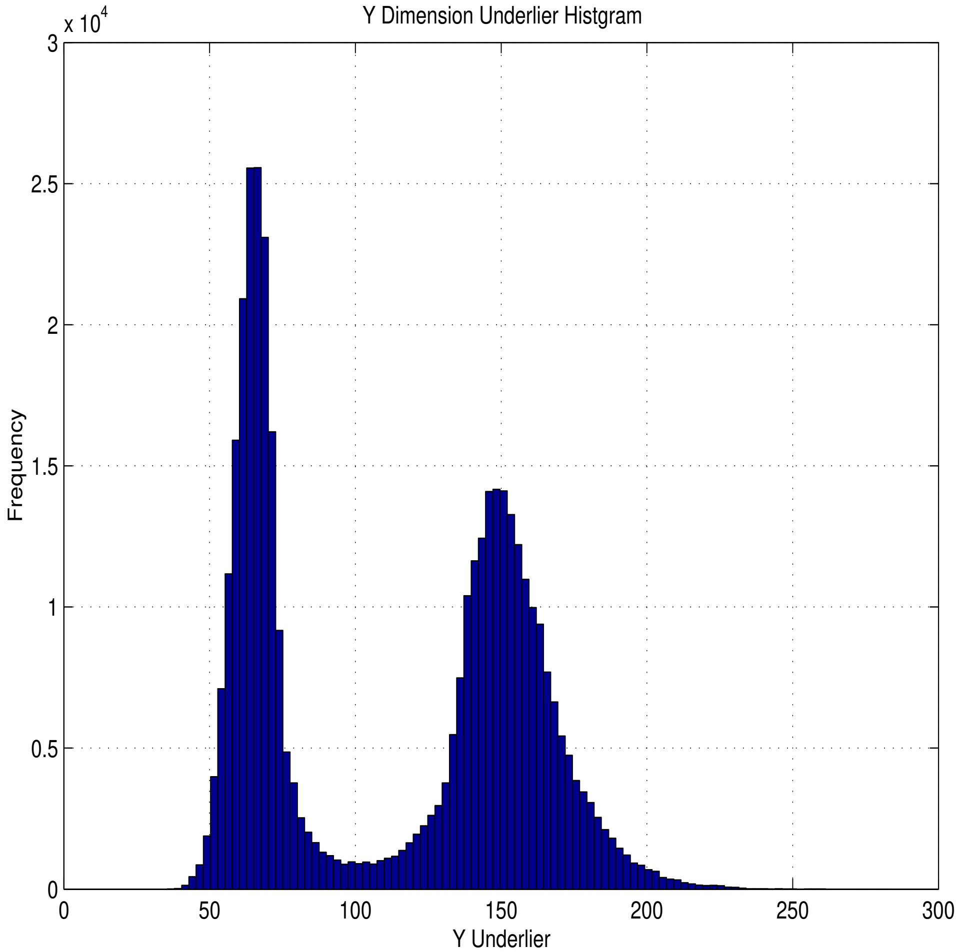

Figure 8.4: (The Histgram for the Y Underlier: Butterfly Correlated with the Diagonal = 0.5) This corresponds to the same portfolio taken up in the preious page. The picture clearly indicates that π_θ is peaked near the VAR region.

Chapter 9

Summary

In this paper, we have demonstrated effective simulation methods for rare event simulation in finance. In particular, we have applied our methods to compute VAR probabilities using both i.i.d. Monte Carlo, and the Markov Chain Monte Carlo (MCMC). For the i.i.d. Monte Carlo, we have presented our new algorithm based on importance sampling combined with non-linear optimization with our new exponential twisted density. The idea is to use a convex combinations of exponential twisted densities along the direction perpendicular to the VAR surface at each local minima β which are found by non-linear optimization. Our experimental results have shown that this method outperforms the Gaussian twist with shifted means. Experimentally variance ratios for our new twist were significantly larger than the Gaussian counterpart for virtually all portfolio by factors ranging from 2 to 30. The effectiveness of this method was shown analytically by the Laplace method (the asymptotic approximation theory) as well (Appendix C). Even from the Large Deviation Theory point of view, we have shown that this method is asymptotically efficient if all local minima are taken into account.

The byproduct of these experiments is the confirmation that all other analytic based importance sampling techniques fail when there are more than one local minimum. This happens for various delta-hedged portfolio as independently confirmed by GHS ([41]). Another discovery of our study is that for portfolios with non-extreme quantile and one local minimum, the combination of the control variate and the importance sampling method outperformed the importance sampling technique alone.

In the later chapters, we have focused on more difficult portfolios which include path-dependent options, and portfolios with many local minima. We proposed new technique based on the Markov Chain Monte Carlo. Our invention is a family of new target densities of the form $\pi_\theta = f \cdot e^{-\theta max(V,v)}$ parameterized by θ which is picked optimally by the Robbins-Monro algorithm ($\theta_{n+1} = \theta_n + \frac{1}{n} sign(v - v)$). The formula of the function resembles that of the Gibbs distribution used in computational physics. Then sample from π_θ using either the Sequential or the Random Scan Symmetric-Random Walk Metropolis-within-Gibbs algorithm. Our contribution is the combined approach of the Metropolis algorithm and the Robbins-Monro algorithm in this fashion. Our experiments have shown the effectiveness of this method over the plain vanilla Monte Carlo methods. Our experiments further demonstrates that the variance for this method is bigger than the i.i.d. based importance sampling method due to the autocorrelation time. Therefore, this method can be used as a last resort when all other i.i.d. based twisting methods fails.

We have provided several proofs showing that our method converges almost surely for the Random Scan Metropolis-within-Gibbs method. Since for the Random Scan Algorithm, the chains is also proven to be geometrically ergodic, the Large Deviation Theory can be used to show that our new MCMC technique is still asymptotically efficient as long as we samples all possible local minima and the quantile is extreme enough. This is so because the effect from the dependent factors decreases exponentially (or geometrically). Thus, the variance for our MCMC chain is bigger than the i.i.d. MC by a factor called the autocorrelation time (which is bigger than 1 but finite). Therefore, the asymptotic efficiency for our method still dominates if the tail is far enough.

The byproduct of these analysis is our proposal for pricing far-out-of-the-money basket options (puts) since our twisted target density optimally samples most near the region of interest for this case. Furthermore, we have proposed a new method for computing VAR probability for portfolios containing path-dependent options (Asian, Lookback, and American options) using the Monte Carlo, and effectively applied our new MCMC techniques to these cases. Our experiments have shown that our method works quite well even if the number of exercise time (or dimension) is increased by more fine gridding.

However, even this method can fail for some extreme cases when our MCMC is trapped in one local minimum. We believe that our method can be improved by possibly another approach. This can be left as a future research topic. Moreover, the autocorrelation time can grow significantly as the dimension becomes larger for some portfolios. Then, the variance ratio can go below 1 (less effective than the regular Monte Carlo method if it is possible to do). Therefore, one has to be aware of the potential shortcoming of this method. The autocorrelation time may be algorithmically controlled by potentially another approach. This too is an interesting topic of research. Nevertheless, if i.i.d. based sampling is impossible, then this algorithm is the best candidate for sampling rare event scenarios. We believe that our studies opened a door for further MCMC application to VAR and computational finance in general.

[APPENDIX A: Semi-Analytic Formula for Non-Linear Portfolio VAR]

A1 Introduction

In this section, an analytic VAR formula for a portfolio containing non-linear securities will be derived. It is based on a work done by Cardenas (1997). This theorem treats general cases where the number of risk factors m differs from the effective dimension d ($m \geq d$).

Thorem 9.0.1 (Analytic VAR Formula for a Non-Linear Portfolio) *Let us be given m risk factors $S(t) = [S_1(t), S_2(t), ...S_m(t)]$, a set of d independent standard normals x associated with each risk factors such that $m \geq d$, and a portfolio $V \equiv V(S(t))$. Let us define v to be the VAR for a given confidence interval p, and let us define P&L (Profit and Loss) as $V_0 - V$. Let us also assume that S satisfies the following process: $dS_t = Adt + B \cdot dW_t$ where $A \in R^{m,1}$, $B \in R^{m,n}$ are constant matrix, and S_t and W_t have dimension of m, n respectively. Then for a given confidence p, v can be computed by reading off the following inverse FFT (Fast Fourier Transform):*

$$\int_{-\infty}^{v-\bar{a}(t)} q(y)dy \approx 1 - p \tag{A.1}$$

where $q(y)$ is the inverse Fourier transform of $\hat{q}(\zeta)$ whose value is expressed as:

$$\hat{q}(\zeta) = \prod_{j=1}^{d} \hat{q}_j(sign(\lambda_j)\zeta)\hat{q}_0(\zeta) \tag{A.2}$$

where

$$\hat{q}_j(\zeta) = \frac{e^{i\zeta\mu_j^2/(1-2i\zeta\sigma_j^2)}}{\sqrt{1-2i\zeta\sigma_j^2}}, \qquad \hat{q}_0(\zeta) = e^{-\frac{1}{2}b(t)^2[1-\sum_{j=1,\lambda_j\neq 0}^{n}\rho(\xi_0,\psi_j)^2]\zeta^2}$$

$$\begin{aligned}
\mu_j &= \frac{1}{2\sqrt{|\lambda_j|}} Q_j \cdot B^T \cdot \nabla[t(\partial_t + A\cdot\nabla)V + V] \\
\sigma_j &= \sqrt{t|\lambda_j|}, \quad b(t) = \sqrt{t}\|B^T \cdot \nabla[t(\partial_t + A\cdot\nabla)V + V]\| \\
\bar{a}(t) &= -t[(\partial_t + A\cdot\nabla)V] - \frac{t^2}{2}[(\partial_t^2 + 2A\cdot\nabla\partial_t)V + A\cdot\nabla(A\cdot\nabla V)] \\
&\quad - \sum_{j=1,\lambda_j\neq 0}^{n} \frac{(b(t)\rho(\xi_0,\psi_j))^2}{4t\lambda_j}
\end{aligned} \tag{A.3}$$

λ_j is the diagonal component of the diagonal matrix Λ found by the Schur decomposition of the matrix $M = B^T \cdot \nabla^2 V \cdot B = Q^T \cdot \Lambda \cdot Q$, and i is a complex number $\sqrt{-1}$

Proof: Let us assume that V is smooth enough to allow the Ito-Taylor expansion as follows:

$$\begin{aligned}
V_0 - V &= -\int_0^t \frac{\partial V(r,S_r)}{\partial r} dr - \int_0^t \nabla V(r,S_r)\cdot dS_r - \frac{1}{2}\sum_{j,k}\int_0^t \frac{\partial^2 V(r,S_r)}{\partial S_r^j \partial S_r^k} d < S_r^j, S_r^k > \\
&\approx -\int_0^t [\partial_r V(0,S_0) + r\partial_r^2 V(0,S_0) + \nabla\partial_r V(0,S_0)\cdot(rA + B\cdot W_r)]dr \\
&\quad -\int_0^t \nabla[V(0,S_0) + r\partial_r V(0,S_0) + \nabla V(0,S_0)\cdot(rA + B\cdot W_r)]\cdot(Adr + B\cdot dW_r)dr \\
&\quad -\frac{1}{2}\sum_{j,k}\int_0^t \partial_{jk}^2 V(0,S_0)(B\cdot B^T)_{jk} dr \\
&= -t[(\partial_t + A\cdot\nabla)V + \sum_{j,k}\frac{(B\cdot B^T)_{jk}}{2}\partial_{jk}V] - \frac{t^2}{2}[(\partial_t^2 + 2A\cdot\nabla\partial_t)V + A\cdot\nabla(A\cdot\nabla V)] \\
&\quad -\nabla[t(\partial_t + A\cdot\nabla)V + V]\cdot B\cdot W_t - \int_0^t W_r^T \cdot M\cdot dW_r \qquad (A.4)
\end{aligned}$$

where, $M = B^T \cdot \nabla^2 V \cdot B$.
Since $\int_0^t W_r^T \cdot M \cdot dW_r = \frac{1}{2} W_t^T \cdot M \cdot W_t - \frac{t}{2} tr(M)$ and $tr(M) = \sum_{j,k}(B \cdot B^T)_{jk}\partial_{jk}V$, the last term of the $(-t)$ part of the above expression $(A.4)$ disappears. Therefore, the loss function can be further simplified to give:

$$\begin{aligned}
V_0 - V &\approx -t[(\partial_t + A\cdot\nabla)V] - \frac{t^2}{2}[(\partial_t^2 + 2A\cdot\nabla\partial_t)V + A\cdot\nabla(A\cdot\nabla V)] \\
&\quad -\nabla[t(\partial_t + A\cdot\nabla)V + V]\cdot B\cdot W_t - \frac{1}{2}W_t^T\cdot M\cdot W_t \qquad (A.5)
\end{aligned}$$

The last term of the expression $(A.5)$ can be further transformed into matrix products containing diagonal matrix Λ by performing Schur decomposition to the matrix $M = B^T \cdot \nabla^2 V \cdot B$ as follows:

$$\begin{aligned}
-\frac{1}{2}W_t^T\cdot M\cdot W_t &= W_t^T\cdot Q^T\cdot\Lambda\cdot Q\cdot W_t = (Q\cdot W_t)^T\cdot\Lambda\cdot(Q\cdot W_t) \\
&= \sum_{j=1}^d (Q_j\cdot W_t)\lambda_j(Q_j\cdot W_t) \qquad (A.6)
\end{aligned}$$

Letting $\psi = \frac{Q\cdot W_t}{\sqrt{t}}$ which is a standard normal vector yields:

$$V_0 - V \approx a(t) + b(t)\xi_0 + t\sum_{j=1}^d \lambda_j \psi_j^2 \qquad (A.7)$$

where $\xi_0 \sim N(0,1)$, $\psi_j \sim N(0,1)$, $j = 1, 2, ...d$.

$$
\begin{aligned}
a(t) &= -t[(\partial_t + A \cdot \nabla)V] - \frac{t^2}{2}[(\partial_t^2 + 2A \cdot \nabla \partial_t)V + A \cdot \nabla(A \cdot \nabla V)] \\
b(t) &= \sqrt{t}||B^T \cdot \nabla[t(\partial_t + A \cdot \nabla)V + V]|| \\
&= \sqrt{t}\sqrt{(B^T \cdot \nabla[t(\partial_t + A \cdot \nabla)V + V])^T \cdot (B^T \cdot \nabla[t(\partial_t + A \cdot \nabla)V + V])}
\end{aligned}
$$

Note that ξ_0 and ψ_j are not independent, even though each ψ_j is independent of each other. Thus, in order to rectify this problem, one need to remove the dependent component due to each ψ_j from ξ_0 by forming an another random number $\bar{\xi}_0$ given by: $\bar{\xi}_0 = \xi_0 - \sum_{j=1}^{d} \rho(\xi_0, \psi_j)\psi_j$. Thus, $V_0 - V$ is given by:

$$
\begin{aligned}
V_0 - V &\approx a(t) + b(t)\bar{\xi}_0 + t\sum_{j=1}^{d}(t\lambda_j\psi_j^2 + b(t)\rho(\xi_0, \psi_j)\psi_j) \\
&= a(t) + b(t)\bar{\xi}_0 + t\sum_{j=1,\lambda_j \neq 0}^{d}(t\lambda_j\psi_j^2 + b(t)\rho(\xi_0, \psi_j)\psi_j) + \sum_{j=1\lambda_j=0}^{d}[b(t)\rho(\xi_0, \psi_j)\psi_j] \\
&= a(t) + b(t)[\xi_0 - \sum_{j=1,\lambda_j \neq 0}^{d} \rho(\xi_0, \psi_j)\psi_j] + \sum_{j=1,\lambda_j \neq 0}^{d}(t\lambda_j\psi_j^2 + b(t)\rho(\xi_0, \psi_j)\psi_j) \\
&= a(t) + b(t)[\xi_0 - \sum_{j=1,\lambda_j \neq 0}^{d} \rho(\xi_0, \psi_j)\psi_j] \\
&\quad +t\sum_{j=1\lambda_j \neq 0}^{d} \lambda_j[(\psi_j + \frac{(b(t)\rho(\xi_0, \psi_j))}{2t\lambda_j})^2 - (\frac{b(t)\rho(\xi_0, \psi_j)}{2t\lambda_j})^2] \qquad (A.8) \\
&= \bar{a}(t) + b(t)\xi + t\sum_{j=1\lambda_j \neq 0}^{d} \lambda_j[(\psi_j + \frac{(b(t)\rho(\xi_0, \psi_j))}{2t\lambda_j})^2] \qquad (A.9) \\
&= \bar{a}(t) + b(t)\xi + t\sum_{j=1}^{d} \lambda_j^2 x_j^2 \qquad (A.10)
\end{aligned}
$$

where $\xi \sim N[0, 1 - \sum_{j=1,\lambda_j \neq 0}^{d} \rho(\xi_0, \psi_j)]$, $x_j \sim N[\frac{b(t)\rho(\xi_0, \psi_j)}{4t\lambda_j}, 1]$.

$$
\bar{a}(t) = a(t) - \sum_{j=1,\lambda_j \neq 0}^{d} \frac{(b(t)\rho(\xi_0, \psi_j))^2}{4t\lambda_j} \qquad (A.11)
$$

Letting $Y_j = t|\lambda_j|x_j^2$ transforms the above expression into:

$$
V_0 - V \approx \bar{a}(t) + b(t)\xi + \sum_{j=1}^{d} sign(Y_j)Y_j \qquad (A.12)
$$

where each $Y_j = t|\lambda_j|x_j^2$ has density:

$$q_j(y) = \begin{cases} \frac{e^{-(y+\mu_j^2)/(2\sigma_j^2)}}{\sigma_j\sqrt{2\pi y}} cosh[\frac{\mu_j\sqrt{y}}{\sigma_j^2}] & , y > 0 \\ 0 & , y \leq 0. \end{cases}$$

$$\mu_j \quad = \quad \frac{1}{2\sqrt{|\lambda_j|}} Q_j \cdot B^T \cdot \nabla[t(\partial_t + A \cdot \nabla)V + V], \;\; \sigma_j = \sqrt{t|\lambda_j|}, \;\; j = 1, 2, ..., d$$

It is also not hard to show that the above density has the following characteristic function (computation can be performed by completing squares for the exponent of the integral formula for the characteristic function) :

$$\bar{q}_j(\zeta) = \frac{i\zeta\mu_j^2/(1 - 2i\zeta\sigma_j^2)}{\sqrt{1 - 2i\zeta\sigma_j^2}}, \quad j = 1, 2, ...d \tag{A.13}$$

where $i = \sqrt{-1}$. The characteristic function of $b(t)\xi$ can be shown as:

$$\bar{q}_0(\zeta) = e^{[-\frac{1}{2}b(t)^2(1 - \sum_{j=1,\lambda_j \neq 0}^{d} \rho(\xi_0,\psi_j)^2)\zeta^2]} \tag{A.14}$$

Thus the convolution product is given by: $\bar{q}(\zeta) = \prod_{j=1}^{d} \bar{q}_j(sign(\lambda_j)\zeta)\bar{q}_0(\zeta)$.

QED.

The above formula $(A.14)$ requires the computation of correlation with respect to ξ_0 and ψ_j, $j = 1, 2, ..., d$. One can derive another formulation for the above theorem. Let us assume that the number of risk factors equals the number of underlying normal variables (That is $m = d$). Then $(A.5)$ can be rewrittes as:

$$V_0 - V \quad \approx \quad a(t) + \delta(t) \cdot B \cdot W_t - \frac{1}{2} W_t^T \cdot M \cdot W_t \tag{A.15}$$

where, $a(t) = -t[(\partial_t + A \cdot \nabla)V] - \frac{t^2}{2}[(\partial_t^2 + 2A \cdot \nabla\partial_t)V + A \cdot \nabla(A \cdot \nabla V)]$, $\delta(t) = -\nabla[t(\partial_t + A \cdot \nabla)V + V]$, and $M = B^T \cdot \nabla^2 V \cdot B$. Let C be such that $B \cdot W_t = \sqrt{(t)}B \cdot x = C \cdot x$ where x is a d dimansional standard normal. Assume also that the correlation matrix for the risk factord as given by ρ. Then the covariance matrix Σ_x for d risk factors can be computed as $\Sigma_x = C \cdot \rho \cdot C^T$. Let us also desine the Cholesky decomposition of Σ_x as $\Sigma_x = \hat{C} \cdot \hat{C}^T$. So $(A.15)$ becomes:

$$V_0 - V \quad \approx \quad a(t) + \delta(t) \cdot C \cdot x + (C \cdot x)^T \cdot (-\frac{1}{2}\nabla^2 V) \cdot (C \cdot x) \tag{A.16}$$

Moreover, performing Schur decomposition for the second order matrix $\hat{C}^T \cdot (-\frac{1}{2}\nabla^2 V) \cdot \hat{C}$ gives $\hat{C}^T \cdot (-\frac{1}{2}\nabla^2 V) \cdot \hat{C} = Q \cdot \Lambda \cdot Q^T$. Then, by letting $C = \hat{C} \cdot Q$,

$(A.16)$ can be further transformed as (This is fine since $C \cdot C^T = \hat{C} \cdot Q \cdot Q^T \cdot \hat{C}^T = \hat{C} \cdot \hat{C}^T = \Sigma_x$) :

$$
\begin{aligned}
V_0 - V &\approx a(t) + \delta(t) \cdot C \cdot x + (\hat{C} \cdot Q \cdot x)^T \cdot (-\frac{1}{2}\nabla^2 V) \cdot (\hat{C} \cdot Q \cdot x) \\
&= a(t) + \delta(t) \cdot C \cdot x + x^T \cdot Q^T \cdot \hat{C}^T \cdot (-\frac{1}{2}\nabla^2 V) \cdot \hat{C} \cdot Q \cdot x) \\
&= a(t) + \delta(t) \cdot \hat{C} \cdot Q \cdot x + x^T \cdot Q^T \cdot Q \cdot \Lambda Q^T \cdot Q \cdot x) \\
&= a(t) + \delta(t) \cdot \hat{C} \cdot Q \cdot x + x^T \cdot \Lambda \cdot x) = a(t) + \sum_{j=1}^{d} (b_j(t) x_j + \lambda_j x_j^2)
\end{aligned}
$$

where, b_j is the j_{th} component of the vector $\delta(t) \cdot \hat{C} \cdot Q$, and λ_j is the j_{th} component of the diagonal element of the diagonal matrix Λ.
Now each $(b_j(t) x_j + \lambda_j x_j^2)$ is independent for all $j = 1, 2, ...d$. Therefore, one can compute the Fourier Transform (Computation can be done by expanding the Fourier Integral and completing squares. So we shall skip the calculation detail here) to get the following:

$$
\hat{q}_j(\zeta) = \frac{1}{\sqrt{(1 + 4\pi i \zeta \lambda_j)}} e^{\frac{-2\pi^2 \zeta^2 b_j^2}{(1+4\pi i \zeta \lambda_j)}} \tag{A.17}
$$

where i is a complex number $\sqrt{-1}$.

One can think of b_j as a generalized delta component, and λ_j as a generalized gamma component. If $\lambda_j = 0$, then the above $(A.17)$ collapses to:

$$
\hat{q}_j(\zeta) = e^{-2\pi^2 \zeta^2 b_j^2} \tag{A.18}
$$

which is the correct Fourier Transform formula for the delta based factor which is purely Gaussian.

Most cases, one only have delta, gamma, and theta. Let us assume that dS can be approximated as : $dS_t = \mu S_t dt + \sigma \cdot S_t \cdot dW_t \approx (\mu - \frac{1}{2}\sigma^2) S_0 dt + \sigma \cdot S_0 \cdot dW_t$ where μ is the drift vectr and σ is the volatility vector with covariance matrix Σ_S and correlation matrix ρ. Then we can perform the schur decomposition of $\hat{C}^T \cdot (-\frac{1}{2}\nabla^2 V) \cdot \hat{C}$ as $\hat{C}^T \cdot (-\frac{1}{2}\nabla^2 V) \cdot \hat{C} = Q \cdot \Lambda \cdot Q^T$ where $\hat{C}$ is the Cholesky decomposition of Σ_S such that $(\sigma \cdot S_0) \cdot \rho \cdot (\sigma \cdot S_0)^T \equiv \Sigma_S = \hat{C} \cdot \hat{C}^T$. Then v be approximated by reading off the following inverse FFT:

$$
\int_{-\infty}^{v - a(t)} q(y) dy \approx 1 - p
$$

where $q(y)$ is the inverse FFT of $\hat{q}(\zeta)$ given by: $\hat{q}(\zeta) = \prod_{j=1}^{d} \hat{q}_j(sign(\lambda_j)\zeta)$, and $\hat{q}_j(\zeta)$ is given by $(A.17)$. $a(t) = -t\partial_t V$, λ_j is the j_{th} component of Λ, $b_j(t)$ is the j_{th} component of $-\nabla V \cdot \hat{C} \cdot Q$, and i is a complex number $\sqrt{-1}$.

[APPENDIX B: The GHS Method]

B1 Introduction

In this section we will describe a method proposed by GHS ([41]). This algorithm create a Gaussian importance sampling twisted density by choosing a mean and covariance matrix using the Taylor approximation of the portfolio V as a guidance. We have used this method both as a comparison purpose, and as an initial starting guess point for our non-linear optimization. We shall describe the algorithm below. Original authors call this Taylor expansion as the'Delta-Gamma' approximation.

Let us assume that the portfolio function V is smooth enough to allow Taylor expanding V as follows:

$$V_0 - V \approx -\frac{\partial V}{\partial t}\Delta t - \delta^T \cdot \Delta S - \frac{1}{2}\Delta S^T \cdot \Gamma \cdot \Delta S = -\Theta \Delta t + Q \quad \text{(B.1)}$$

where V_o is the initial portfolio value, and

$$\delta = [\frac{\partial V}{\partial S_1}, \frac{\partial V}{\partial S_2}, ..., \frac{\partial V}{\partial S_d}], \text{ and } \Gamma = \begin{pmatrix} \frac{\partial^2 V}{\partial S_1 \partial S_1} & \cdots & \frac{\partial^2 V}{\partial S_1 \partial S_d} \\ \cdots\cdots\cdots\cdots\cdots\cdots & \cdots & \cdots \\ \frac{\partial^2 V}{\partial S_d \partial S_1} & \cdots & \frac{\partial^2 V}{\partial S_d \partial S_d} \end{pmatrix} \quad \text{(B.2)}$$

with a correlation matrix Σ_x. The GHS method calls for Cholesky decomposing this matrix to get $\hat{C}$ so that $\hat{C}^T \cdot \hat{C} = \Sigma_x$. The GHS method then calls for performing the Schur decomposition of Σ_x so that $\hat{C}^T \cdot (-\frac{1}{2}\Gamma) \cdot \hat{C} = U \cdot \Lambda \cdot U^T$ where U is an unitary matrix. U is an orthogonal matrix whose columns are the corresponding unit eigenvectors of $\hat{C}^T \cdot \hat{C}$, and Λ is a diagonal matrix for the eigenvalues of $\hat{C}^T \cdot (-\frac{1}{2}\Gamma) \cdot \hat{C}$. By letting $C = \hat{C} \cdot U$, $b = -\delta^T \cdot C$, and $\Delta S = C \cdot x$ where x is a vector of d independent standard normals (with mean zero and variance one), we have:

$$\begin{aligned} V_0 - V &\approx -\frac{\partial V}{\partial t}\Delta t + Q = -\Theta \Delta t - \delta^T \cdot \Delta S - \frac{1}{2}\Delta S^T \cdot \Gamma \cdot \Delta S \\ &= -\Theta \Delta t - \delta^T \cdot C \cdot x - (\hat{C} \cdot U \cdot x)^T \cdot (-\frac{1}{2}\Gamma) \cdot (\hat{C} \cdot U \cdot x) \\ &= -\Theta \Delta t + b^T \cdot x + x^T \cdot \Lambda \cdot x \equiv -\Theta \Delta t + \sum_{i=1}^{d}(b_i x_i + \lambda_i x_i^2) \quad \text{(B.3)} \end{aligned}$$

where Λ is an diagonal matrix which can be rearranged as:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & ... & 0 & 0 \\ 0 & \lambda_2 & ... & 0 & 0 \\ .... & ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_{d-1} & 0 \\ 0 & 0 & ... & 0 & \lambda_d \end{pmatrix}, \lambda_1 \geq \lambda_2 \geq ..., \geq \lambda_d \tag{B.4}$$

and the set $\{\lambda_i\}$ are the eigenvalues of $-\frac{1}{2}\Gamma \cdot \Sigma_x$ so that $\Lambda = -\frac{1}{2}C^T \cdot \Gamma \cdot C$, and $C \cdot C^T = \Sigma_x$. Since $P(V_0 - V \geq v) \approx P(Q \geq -\Theta\Delta t + v)$, the importance likelihood ratio function suggested by GHS is the following formula:

$$m(x) \quad = \quad e^{-\theta Q + \phi(\theta)} \tag{B.5}$$

where θ is any positive real constant so that $\theta > 0$, and $\phi(\theta)$ is a the logarithm of the moment generating function of Q given by:

$$\phi(\theta) \quad = \quad \frac{1}{2}\sum_{i=1}^{d}[\frac{(\theta b_i)^2}{(1-2\theta\lambda_i)} - Log(1-2\theta\lambda_i)] \tag{B.6}$$

GHS suggested to find θ so that for a given v, the following is true:

$$\phi(\theta)^{'} \quad = \quad \frac{d\phi(\theta)}{d\theta} = \sum_{i=1}^{d}[\frac{\theta b_i^2(1-\theta\lambda_i)}{(1-2\theta\lambda_i)^2} + \frac{\lambda_i}{(1-2\theta\lambda_i)}] = v \tag{B.7}$$

This can be done by performing the Newton's method to compute its root θ. Once the root is found, then the following twisted density is created:

$$\pi_\theta(x) \quad = \quad \frac{1}{(2\pi)^{d/2}|\Sigma_\theta|^{1/2}}e^{-\frac{1}{2}(x-\mu_\theta^T \cdot \Sigma_\theta^{-1} \cdot (x-\mu_\theta)} \tag{B.8}$$

where $\Sigma_\theta = (1-2\theta\Lambda)^{-1}$, and $\mu_\theta = \theta\Sigma_\theta \cdot b$ which can be expressed componentwide as:

$$\sigma_i^2(\theta) = \frac{1}{(1-2\theta\lambda_i)}, \quad \mu_i(\theta) = \frac{\theta b_i}{(1-2\theta\lambda_i)} \tag{B.9}$$

The next short theorem gives the justification for the above formulae.

Thorem 9.0.2 (Analytic Twisted Density Formula)
Let us be given all the notations as in the previous paragraphs. Then the Gaussian twist given by (B.8) satisfies $E_{\mu_\theta,\Sigma_\theta}(Q) = v$.

Proof: Given the probability densities, $\pi_\theta(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_\theta|^{1/2}}e^{-\frac{1}{2}(x-\mu_\theta)^T \cdot \Sigma_\theta^{-1} \cdot (x-\theta)}$, and $f(x) = \frac{1}{(2\pi)^{d/2}}e^{-\frac{1}{2}x^T \cdot x}$, we can express $P(V_0 - V \leq v)$ as:

$$P(V_0 - V \leq v) = E(I(V_0 - V \leq v)) = E_{\mu_\theta,\Sigma_\theta}(m(x)I(V_0 - V \leq v)) \tag{B.10}$$

where $m(x) = \frac{f(x)}{\pi_\theta(x)} = |\Sigma_\theta|^{1/2} e^{-\frac{1}{2}x^T\cdot x + \frac{1}{2}(x-\mu_\theta)^T\cdot\Sigma_\theta^{-1}\cdot(x-\mu_\theta)} \equiv |\Sigma_\theta|^{1/2} e^{\alpha}$.
α can be expanded as:

$$\begin{aligned}
\alpha &= -\frac{1}{2}x^T\cdot x + \frac{1}{2}x^T\cdot\Sigma_\theta^{-1}\cdot x - \frac{1}{2}x^T\cdot\Sigma_\theta^{-1}\cdot\mu_\theta - \frac{1}{2}\mu_\theta^T\cdot\Sigma_\theta^{-1}\cdot x + \frac{1}{2}\mu_\theta^T\cdot\Sigma_\theta^{-1}\cdot\mu_\theta \\
&= -\frac{1}{2}x^T\cdot x + \frac{1}{2}x^T\cdot(I-2\theta\Lambda)\cdot x - \frac{1}{2}x^T\cdot\Sigma_\theta^{-1}\cdot(\theta\Sigma_\theta\cdot b) \\
&\quad -\frac{1}{2}(\theta\Sigma_\theta\cdot b)^T\cdot\Sigma_\theta^{-1}\cdot x + \frac{1}{2}\mu_\theta^T\cdot\Sigma_\theta^{-1}\cdot\mu_\theta \\
&= -\frac{1}{2}x^T\cdot x + \frac{1}{2}x^T\cdot x - \theta(x^T\cdot\Lambda\cdot x) \\
&\quad -\frac{\theta}{2}x^T\cdot b - \frac{\theta}{2}b^T\cdot x + \frac{1}{2}(\theta\Sigma_\theta\cdot b)^T\cdot\Sigma_\theta^{-1}\cdot(\theta\Sigma_\theta\cdot b) \\
&= -\theta(x^T\cdot\Lambda\cdot x + b^T\cdot x) + \frac{\theta}{2}(\theta\Sigma_\theta\cdot b)^T\cdot b\frac{1}{2}x^T\cdot x + \frac{1}{2}x^T\cdot x - \theta(x^T\cdot\Lambda\cdot x) \\
&= -\theta\, Q + R
\end{aligned}$$

where $R = \frac{1}{2}\theta^2 b^T\cdot(I-2\theta\Lambda)^{-1}\cdot b = \frac{1}{2}\sum_{i=1}^{d}[\frac{(\theta b_i)^2}{(1-2\theta\lambda_i)}]$. Therefore, $m(x)$ can be expressed as:

$$m(x) = |\Sigma_\theta|^{1/2}e^{\alpha} = e^{-\theta Q + R + \frac{1}{2}Log(|\Sigma_\theta|)} \tag{B.11}$$

Since $|\Sigma_\theta| = |(I-2\theta\lambda)^{-1}| = \prod_{i=1}^{d}(1-2\theta\lambda_i)^{-1}$, we can express the above expression as follows:

$$m(x) = |\Sigma_\theta|^{1/2}e^{\alpha} = e^{-\theta Q + \frac{1}{2}\sum_{i=1}^{d}[\frac{(\theta b_i)^2}{(1-2\theta\lambda_i)} - Log(1-2\theta\lambda_i)]} = e^{-\theta Q + \phi(\theta)}$$

which is basically $(B.5)$. Taking an expectation of $e^{\theta Q}$ under the original measure, and using the relation above gives::

$$E[e^{\theta Q}] = E_{\mu_\theta,\Sigma_\theta}[m(x)e^{\theta Q}] = E_{\mu_\theta,\Sigma_\theta}[e^{\phi(\theta)}] = e^{\phi(\theta)} \tag{B.12}$$

Therefore, one obtains $\phi(\theta) = Log[E[e^{\theta Q}]]$ which is basically a logmoment generating function. Using the relation $m(x) = e^{-\theta Q}e^{\phi(\theta)}$, the derivative of $\phi(\theta)$ with respect to θ is given by:

$$\begin{aligned}
\frac{d\phi(\theta)}{d\theta} &= \frac{E[Qe^{\theta Q}]}{E[e^{\theta Q}]} = \frac{E_{\mu_\theta,\Sigma_\theta}[m(x)Qe^{\theta Q}]}{E_{\mu_\theta,\Sigma_\theta}[m(x)e^{\theta Q}]} = \frac{E_{\mu_\theta,\Sigma_\theta}[e^{-\theta Q}e^{\phi(\theta)}Qe^{\theta Q}]}{E_{\mu_\theta,\Sigma_\theta}[e^{-\theta Q}e^{\phi(\theta)}e^{\theta Q}]} \\
&= \frac{e^{\phi(\theta)}E_{\mu_\theta,\Sigma_\theta}[Q]}{e^{\phi(\theta)}} = E_{\mu_\theta,\Sigma_\theta}[Q]
\end{aligned}$$

Therefore, when θ is chosen to satisfy $\frac{d\phi(\theta)}{d\theta} = v$, the theorem is proved since $\frac{d\phi(\theta)}{d\theta} = E_{\mu_\theta,\Sigma_\theta}[Q]$.

QED.

APPENDIX C: Analytic Efficiency Formula for Importance Sampling

C1 Introduction

In this section, the effectiveness of both Gaussian and exponential importance sampling twisted densities (introduced in chapter 3) will be shown for computing VAR probability. This will be done by computing 'efficiency' analytically. We shall define the term 'efficiency' of method 2 over method 1 as the variance ratio of these methods (i.e., $\frac{\sigma_1^2}{\sigma_2^2}$ where σ_1^2 is the variance for the method 1, and σ_2^2 is the variance for the method 2). If efficiency is greater than one, then the method 2 is said to be more efficient than the method 1. At first, we will treat the VAR region B_v as a d-dimensional sphere and derive the efficiency formula for both Gaussian and exponential twists over the regular Monte Carlo method. Then we extend the results to more general convex set B_v for the rare event scenarios to show analytically that the exponential twist outperforms the Gaussian twist significantly. The analysis will be done using the Laplace method for rare event simulation. This section confirms that for rare event simulations, the tail of distributions for V behaves like an exponential distribution.

C2 (B_v: d-Dimensional Sphere)

Let us also assume that our portfolio V contains n securities V_i, $i = 1, 2, ...d$ so that $V = \sum_i^d V_i$. We shall follow the same notations as in the Appendix B. Let us also assume that each V_i is smooth enough so that we can Taylor expand V as:

$$V - V_o \quad \approx \quad \equiv -\Theta\Delta t + \sum_{i=1}^{d}(b_i x_i + \lambda_i x_i^2)$$

where b_i, and λ_i are the generalized delta and gamma component respectively as in [1]$(B.3)$. Furthermore, without the loss of generality, let us assume that $-\Theta\Delta t = 0$ (Otherwise, one could always translate the $V - V_o$ by this constant). Therefore, the VAR probability is given by:

$$P(V - V_o \leq -v) \quad \approx \quad P(\sum_{i=1}^{d} b_i x_i + \sum_{i=1}^{d} \lambda_i x_i^2 \leq -v) \tag{C.1}$$

[1]In Appendix B, $V_o - V$ is Taylor expanded. Without the loss of generality, we shall use the same notations for the Taylor expansion of $V - V_o$.

where [2] $-v$ is a given VAR. Let us further assume that $b_i = b$ and $\lambda_i = \lambda > 0$ for all i (Other cases are similar). Dividing the expression for both sides of the inequality inside the above probability expression $(C.1)$ by λ, the probability expression becomes:

$$p \equiv P(V - V_o \leq -v) \quad \approx \quad P(\sum_{i=1}^{d} a x_i + \sum_{i=1}^{d} x_i^2 \leq -\bar{V}) \qquad \text{(C.2)}$$

where $\bar{V} = \frac{v}{\lambda}$, and $a = \frac{b}{\lambda}$. This is the probability of underlying process falling inside a sphere in a d dimensional space. If one drops the assumption that the coefficients b_i, λ_i are the same constant, then the regions of interest changes from a sphere to a n dimensional ellipse or parabola.

Computing this quantity via Monte Carlo simulation can be intuitively improved by shifting the mean from origin to a point where it is the shortest distance from the origin to the sphere in question. Finding this point gives the most dominant direction which influences the probability computation most. Let us call this point β. This should be the local minima for this model problem. There are two candidates for the importance sampling twisted density - the Gaussian distribution $\pi_g(x)$, and the exponential distribution $\pi_e(x)$ given as follows:

$$\pi_g(x) \quad \equiv \quad \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+\beta)\cdot(x+\beta)^T}{2}} \qquad \text{(C.3)}$$

$$\pi_e(x) \quad \equiv \quad \beta e^{\beta\cdot(x+\beta)}\ , \ \ x \leq \beta. \qquad \text{(C.4)}$$

Since B_v is a d-dimensional sphere, one can rotate the region to one of the coordinate, say x_1 (Figure 2). Thus, β for $\pi_e(x)$ is a scalar (> 0), and $\pi_e(x)$ is constructed at $(-\beta, 0, 0, ...0)$ away from the origin. With a slight abuse of notation, β for $\pi_g(x)$ is meant to be a vector $(-\beta, 0, 0, ...0)$. Obviously, the new density with the shifted mean would sample more near the region of interest, so the variance is expected to be smaller. In the next section, efficiency formulae (defined as the variance ratio between 2 different method) will be analytically computed, and explored in full detail. We shall provide a simple theorem concerning variance ratios for both techniques. It is based on computing integrals by a change of coordinates and an application of asymptotic formula called the Laplace method.

Thorem 9.0.3 (Analytic Formula for the Monte Carlo Efficiency) *Suppose we have the assumptions as given in the previous paragraph. Let us define the Monte Carlo efficiency as $\frac{\sigma^2}{\sigma_M^2}$ where σ^2 is the variance for the probability in (C.2) using the regular Monte Carlo method, and σ_M^2 is the variance for the same probability using importance sampling (either Gaussian or exponential twist). Let us define the $Efficiency = \frac{\sigma^2}{\sigma_M^2} = \frac{p-p^2}{M_2-p^2} = \frac{1-p}{\frac{M_2}{p}-p}$, and*

[2]In our main paper, v was denoted as VAR. Here, we shall treat v as a positive number, so the VAR is -v. The reason will be made clear later.

$\beta = \frac{\sqrt{d}a}{2} - \sqrt{\frac{da^2}{4} - \bar{V}}$ *where* $\bar{V} \in [0, \frac{da^2}{4}]$. *Then* $\frac{M_2}{p}$ *and* p *are given by:*

$$\frac{M_2}{p} \approx \begin{cases} \frac{e^{-\frac{\beta^2}{2}}}{2} \frac{[1-\frac{1}{4\beta^2}+\frac{3}{8\beta^4}]}{[1-\frac{1}{\beta^2}+\frac{3}{\beta^4}]} [\frac{\frac{\sqrt{d}a}{2}}{\sqrt{\frac{da^2}{4}-\bar{V}+2\beta}}]^{(\frac{d-1}{2})} & \textit{if Gaussian} \\ \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta} \frac{[1-\frac{2}{\beta^2}+\frac{12}{\beta^4}]}{[1-\frac{1}{\beta^2}+\frac{3}{\beta^4}]} & \textit{if Exponential.} \end{cases}$$

$$p \approx \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta}[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}] \frac{1}{\{1 + \frac{\beta}{\sqrt{\frac{da^2}{4}-\bar{V}}}\}^{\frac{d-1}{2}}}$$

Furthermore, as β *becomes larger, we have:*

$$Efficiency \sim \begin{cases} Const \times e^{\frac{\beta^2}{2}} & \textit{if Gaussian} \\ Const \times \beta e^{\frac{\beta^2}{2}} & \textit{if Exponential.} \end{cases}$$

NOTE: The above expression is an asymptotic approximation. Therefore if β is large, but not large enough, it is best to use only the first 2 terms inside the bracket in the expressions above to get:

$$\frac{M_2}{p} \approx \begin{cases} \frac{e^{-\frac{\beta^2}{2}}}{2} \frac{[1-\frac{1}{4\beta^2}]}{[1-\frac{1}{\beta^2}]} [\frac{\frac{\sqrt{d}a}{2}}{\sqrt{\frac{da^2}{4}-\bar{V}+2\beta}}]^{(\frac{d-1}{2})} & \text{if Gaussian} \\ \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta} \frac{[1-\frac{2}{\beta^2}]}{[1-\frac{1}{\beta^2}]} & \text{if Exponential.} \end{cases}$$

$$p \approx \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta}[1 - \frac{1}{\beta^2}] \frac{1}{\{1 + \frac{\beta}{\sqrt{\frac{da^2}{4}-\bar{V}}}\}^{\frac{d-1}{2}}}$$

Proof: Transforming the probability in question gives:

$$\begin{aligned} p &= P(\sum_{i=1}^{d} ax_i + x_i^2 \leq -\bar{V}) = P(\sum_{i=1}^{d}(x_i + \frac{a}{2})^2 \leq \frac{da^2}{4} - \bar{V}) \\ &= P((x_1 + \frac{\sqrt{d}a}{2})^2 + \sum_{i>1} x_i^2 \leq \frac{da^2}{4} - \bar{V}) \end{aligned} \tag{C.5}$$

The above step in $(C.5)$ is justified, since the probability of a region occupying a particular space in in d dimensional standard normal distribution space is invariant under rotation (See Figure 2). Furthermore, let us define $\bar{x}$, β (which is a local minima for this VAR region) as follows:

$$\bar{x} = \frac{\sqrt{d}a}{2} - \sqrt{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2}\,, \quad \beta = \frac{\sqrt{d}a}{2} - \sqrt{\frac{da^2}{4} - \bar{V}}. \tag{C.6}$$

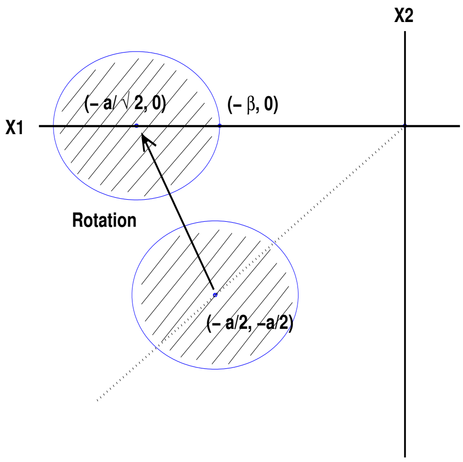

Figure 9.1: 2 Dimensional Case: Since the probability density is a Gaussian centered at the origin, the probability of the two shaded regions are the same. Note that size of β tells how much the new density is shifted away from the origin.

Then p can be expressed as:

$$
\begin{aligned}
p &= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{x_i \in R, i>1, \{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2 \geq 0\}} .. \int_{x_1 < -\bar{x}} e^{-\sum_{i=1}^{d} x_i^2/2} \prod_{i=1}^{d} dx_i \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{x_i \in R, i>1, \{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2 \geq 0\}} .. \int_{x_1 > \bar{x}} e^{-\sum_{i=1}^{d} x_i^2/2} \prod_{i=1}^{d} dx_i
\end{aligned}
$$

In general, for a large value y, one can approximate a Gaussian type integral with simple asymptotic formula as follows by applying the Laplace method [Copson, p.39], one can have the following approximation:

$$
\begin{aligned}
\int_{x>y} e^{-x^2/2} dx &= \int_{x>y} e^{-(y+x-y)^2/2} dx \\
&= e^{-y^2/2} \int_{x>y} e^{-(x-y)y-(x-y)^2/2} dx \\
&= e^{-y^2/2} \int_{r>0} e^{-ry-r^2/2} dr \qquad \text{(C.7)}
\end{aligned}
$$

By using the Taylor expansion for $e^{-r^2/2} \approx 1 - r^2/2 + r^4/8$, we have :

$$
\approx \quad e^{-y^2/2} \int_{r>0} e^{-ry} [1 - r^2/2 + r^4/8] dr \qquad \text{(C.8)}
$$

(Note that the error terms above is of order $o(y)$, this is the crusial part of the asymptotic approximation in that it is accurate for a very large y). By successively performing integrations by parts yields:

$$= \quad e^{-y^2/2}[1/y - 1/y^3 + 3/y^5] = e^{-y^2/2}1/y[1 - 1/y^2 + 3/y^4]$$

Thus, using the above result in the x_1 integral which is the most dominant integral, we have the following approximation:

$$p \quad \approx \quad \int_{x_i \in R, i>1, \{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2 \geq 0\}} \frac{1}{(2\pi)^{\frac{d}{2}}\bar{x}}[1 - \frac{1}{\bar{x}^2} + \frac{3}{\bar{x}^4}]e^{-\bar{x}^2/2 - \sum_{i>1} x_i^2} \prod_{i>1}^{d} dx_i$$

This integral can not computed analytically. However, using them fact that β and $\bar{x}$ is very close enough, $\bar{x}$ in the denominator can be replaced by β. This can be seen by the following approximation for $\bar{x}$:

$$\begin{aligned} \bar{x} \quad &= \quad \frac{\sqrt{d}a}{2} - \sqrt{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2} = \frac{\sqrt{d}a}{2} - \sqrt{\frac{da^2}{4} - \bar{V}}[1 - \frac{\sum_{i>1} x_i^2}{\frac{da^2}{4} - \bar{V}}]^{\frac{1}{2}} \\ &\approx \quad \frac{\sqrt{d}a}{2} - \sqrt{\frac{da^2}{4} - \bar{V}}[1 - \frac{1}{2}\frac{\sum_{i>1} x_i^2}{\frac{da^2}{4} - \bar{V}}] = \beta + \frac{\sum_{i>1} x_i^2}{2\sqrt{\frac{da^2}{4} - \bar{V}}} \qquad \text{(C.9)} \end{aligned}$$

Note that since $2\sqrt{\frac{da^2}{4} - \bar{V}}$ term is very large, and the variation for the non essential coordinates are small, $\sum_{i>1} x_i^2$ term is very small. Thus the entire expression $\frac{\sum_{i>1} x_i^2}{2\sqrt{\frac{da^2}{4} - \bar{V}}}$ is small compared to the order of magnitude of β. Thus we can use the relation $\bar{x} \sim \beta$ in denominator for the probability. Therefore, the probability can be approximated as:

$$p \quad \approx \quad \frac{1}{(2\pi)^{\frac{d}{2}}\beta}[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}] \int_{x_i \in R, i>1, \{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2 \geq 0\}} e^{-\bar{x}^2/2 - \sum_{i>1} x_i^2} \prod_{i>1}^{d} dx_i$$

Furthermore, $\bar{x}^2$ can be approximated by keeping only up to the second power of $\frac{\sum_{i>1} x_i^2}{\sqrt{\frac{da^2}{4} - \bar{V}}}$. Therefore, we have $\bar{x}^2 \approx \beta^2 + \frac{\beta \sum_{i>1} x_i^2}{\sqrt{\frac{da^2}{4} - \bar{V}}}$, and

$$p \quad \approx \quad \frac{e^{-\frac{\beta^2}{2}}}{(2\pi)^{\frac{d}{2}}\beta}[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}] \int_{x_i \in R, i>1, \{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2 \geq 0\}} e^{-\alpha/2 \sum_{i>1} x_i^2} \prod_{i>1}^{d} dx_i$$

where α is given by $\alpha = 1 + \frac{\beta}{\sqrt{\frac{da^2}{4} - \bar{V}}}$. The last remaining integrals are multi-dimensional integrals in $d - 1$ dimensions. This can be analytically computed exactly by transforming into multi-dimensional spherical coordinates, and gives the followings:

$$p \quad \approx \quad \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta}[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}] \frac{1}{(2\pi)^{\frac{d-1}{2}}} \int_{x_i \in R, i>1, \{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2 \geq 0\}}$$

$$
\begin{aligned}
& e^{-\alpha/2\sum_{i>1} x_i^2} \prod_{i>1}^{d} dx_i \\
= \; & \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta}[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}]\frac{1}{\{\alpha\}^{\frac{d-1}{2}}}\{1-\varepsilon\} \\
= \; & \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta}[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}]\frac{1}{\{1+\frac{\beta}{\sqrt{\frac{da^2}{4}-\bar{V}}}\}^{\frac{d-1}{2}}}\{1-\varepsilon\}. \qquad \text{(C.10)}
\end{aligned}
$$

where ε is expressed as:

$$
\varepsilon = \begin{cases} \frac{2}{\sqrt{\pi}}\int_{-\infty}^{-\sqrt{\alpha}\rho} e^{-\frac{y^2}{2}}dy + \sqrt{\frac{2}{\pi}}\alpha^{(\frac{d}{2})}\{\sum_{i=1}^{[\frac{d-1}{2}]}\frac{(\frac{d-i}{2})!2^{(\frac{d-i}{2})}}{\alpha^i(d-i)!}\rho^{(d-2i)}\}e^{-\frac{\alpha\rho^2}{2}} & \text{if } d = odd \\ \alpha^{(\frac{d}{2})}\{1+\sum_{i=1}^{[\frac{d-1}{2}]}\frac{1}{\alpha^i(\frac{d}{2}-i)!2^{(\frac{d}{2}-i)}}\rho^{(d-2i)}\}e^{-\frac{\alpha\rho^2}{2}} & \text{if } d = even. \end{cases}
$$

and, $\rho = \sqrt{\frac{da^2}{4} - \bar{V}}$. Furthermore, since ε is very small and therefore can be dropped from the expression for p, thus we have:

$$
p \approx \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta}[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}]\frac{1}{\{1+\frac{\beta}{\sqrt{\frac{da^2}{4}-\bar{V}}}\}^{\frac{d-1}{2}}}. \qquad \text{(C.11)}
$$

Note that this is basically the same as integrating the remaining $n-1$ integrals for the entire $R^{(d-1)}$ space, rather than computing for the domain given by: $\{x_i \in R, i > 1, \{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2 \geq 0\}\}$.

Next we will compute the expression for the second moment using the Gaussian twisted density with the shifted mean β. Since $\pi_g(x) \equiv \frac{1}{\sqrt{2\pi}}e^{-\frac{(x+\beta)^2}{2}}$, and $f(x) \equiv \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, the second moment $M_2(g)$ for the probability under the Gaussian twist is given by:

$$
\begin{aligned}
M_2(g) = \; & \int_{x_i \in R, i>1, \{\frac{da^2}{4}-\bar{V}-\sum_{i>1} x_i^2/2 \geq 0\}} \cdots \int_{x_1<-\bar{x}} \\
& (\frac{f(x_1)}{\pi_g(x_1)})^2 \pi_g(x_1)dx_1 \prod_{i>1} f(x_i)dx_i \\
= \; & \frac{1}{(2\pi)^{\frac{d}{2}}}\int_{x_i\in\Theta_i} \cdots \int_{x_1<-\bar{x}} (\frac{e^{-\frac{x_1^2}{2}}}{e^{-\frac{(x_1+\beta)^2}{2}}})^2 e^{-\frac{(x_1+\beta)^2}{2}} e^{-\frac{1}{2}\sum_{i>1} x_i^2} \prod_{i\geq 1} dx_i \\
= \; & \frac{1}{(2\pi)^{\frac{d}{2}}}\int_{x_i\in\Theta_i} \cdots \int_{x_1<-\bar{x}} e^{-x_1^2+\frac{(x_1^2+\beta)^2}{2}-\sum_{i>1} x_i^2/2} \prod_{i\geq 1} dx_i \\
= \; & \frac{1}{(2\pi)^{\frac{d}{2}}}\int_{x_i\in\Theta_i} \cdots \int_{x_1<-\bar{x}} e^{-x_1^2+x_1\beta+\frac{\beta^2}{2}-\sum_{i>1} x_i^2/2} \prod_{i\geq 1} dx_i
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{e^{\beta^2}}{(2\pi)^{\frac{d}{2}}}\int_{x_i\in\Theta_i}\cdots\int_{x_1<-\bar{x}} e^{-\frac{(x_1-\beta)^2}{2}-\sum_{i>1}x_i^2/2}\prod_{i\geq 1}dx_i \\
&= \frac{e^{\beta^2}}{(2\pi)^{\frac{d}{2}}}\int_{x_i\in\Theta_i}\cdots\int_{x_1<-\bar{x}-\beta} e^{-\frac{y^2}{2}-\sum_{i>1}x_i^2/2}dy\prod_{i>1}dx_i
\end{aligned}
$$

where Θ_i is the domain $\Theta_i \equiv \{x_i : x_i \in Re, i > 1, \{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2 \geq 0\}$. Applying the Laplace method as before into the above moment expression, and simplifying it further yields:

$$
\begin{aligned}
M_2(g) \approx & \int\cdots\int_{x_i\in\Theta_i}\frac{e^{\beta^2}}{(2\pi)^{\frac{d}{2}}(\bar{x}+\beta)}[1-\frac{1}{(\bar{x}+\beta)^2}+\frac{3}{(\bar{x}+\beta)^4}] \\
& e^{-\frac{(\bar{x}+\beta)^2}{2}-\sum_{i>1}x_i^2/2}\prod_{i>1}dx_i
\end{aligned}
$$

By looking into the approximation formula for $\bar{x}$ and $\bar{x}^2$ in terms of β which were given in the previous section, one can easily realize that $\bar{x}$ can be approximated reasonably by β for the denominator of the above moment expression. Thus:

$$
M_2(g) \approx \frac{e^{\beta^2}}{(2\pi)^{\frac{d}{2}}(2\beta)}[1-\frac{1}{(2\beta)^2}+\frac{3}{(2\beta)^4}]\int\cdots\int_{x_i\in\Theta_i} e^{-\frac{(\bar{x}+\beta)^2}{2}-\sum_{i>1}x_i^2/2}\prod_{i>1}dx_i
$$

Expanding the expression inside the exponent by putting the approximation formulas for $\bar{x}$ and $\bar{x}^2$ gives:

$$
\begin{aligned}
&\approx \frac{e^{\beta^2}}{(2\pi)^{\frac{d}{2}}(2\beta)}[1-\frac{1}{(2\beta)^2}+\frac{3}{(2\beta)^4}]\int\cdots\int_{x_i\in\Theta_i} e^{-\frac{\bar{x}^2}{2}-\bar{x}\beta-\frac{\beta^2}{2}-\sum_{i>1}x_i^2/2}\prod_{i>1}dx_i \\
&= \frac{e^{-\beta^2}}{(2\pi)^{\frac{d}{2}}(2\beta)}[1-\frac{1}{(2\beta)^2}+\frac{3}{(2\beta)^4}]\int\cdots\int_{x_i\in\Theta_i} e^{-\frac{\alpha_G}{2}\sum_{i>1}x_i^2}\prod_{i>1}dx_i
\end{aligned}
$$

where α_G is given by $\alpha_G = 1 + \frac{2\beta}{\sqrt{\frac{da^2}{4}-\bar{V}}}$. This is the familiar expression which can be analytically computed in the multi-dimensional spherical coordinates to give:

$$
\begin{aligned}
M_2(g) &= \frac{e^{-\beta^2}}{\sqrt{2\pi}(2\beta)}[1-\frac{1}{(2\beta)^2}+\frac{3}{(2\beta)^4}](\frac{1}{\alpha_G})^{(\frac{d-1}{2})}\{1-\varepsilon_G\} \\
&\approx \frac{e^{-\beta^2}}{\sqrt{2\pi}(2\beta)}[1-\frac{1}{(2\beta)^2}+\frac{3}{(2\beta)^4}](\frac{1}{\alpha_G})^{(\frac{d-1}{2})} \\
&= \frac{e^{-\beta^2}}{\sqrt{2\pi}(2\beta)}[1-\frac{1}{(2\beta)^2}+\frac{3}{(2\beta)^4}](\frac{\sqrt{\frac{da^2}{4}-\bar{V}}}{\sqrt{\frac{da^2}{4}-\bar{V}}+2\beta})^{(\frac{d-1}{2})} \qquad (C.12)
\end{aligned}
$$

where

$$\varepsilon = \begin{cases} \frac{2}{\sqrt{\pi}} \int_{-\infty}^{-\sqrt{\alpha_G}\rho} e^{-\frac{y^2}{2}} dy + \sqrt{\frac{2}{\pi}} \alpha_G^{(\frac{d}{2})} \{\sum_{i=1}^{[\frac{d-1}{2}]} \frac{(\frac{d-i}{2})! 2^{(\frac{d-i}{2})}}{\alpha_G^i (d-i)!} \rho^{(d-2i)}\} e^{-\frac{\alpha_G \rho^2}{2}} & \text{if } d = odd \\ \alpha^{(\frac{d}{2})} \{1 + \sum_{i=1}^{[\frac{d-1}{2}]} \frac{1}{\alpha_G^i (\frac{d}{2}-i)! 2^{(\frac{d}{2}-i)}} \rho^{(d-2i)}\} e^{-\frac{\alpha_G \rho^2}{2}} & \text{if } d = even. \end{cases}$$

The efficiency of two different Monte Carlo methods is defined as the variance ratio of these methods. Thus by using $(C.11)$ and $(C.12)$, one can ge the following efficiency formula for the Gaussian case.

$$Efficiency = \frac{\sigma^2}{\sigma_M^2} = \frac{p - p^2}{M_2(g) - p^2} = \frac{1 - p}{\frac{M_2(g)}{p} - p}$$

where

$$\frac{M_2(g)}{p} = \frac{e^{-\frac{\beta^2}{2}}}{2} \frac{[1 - \frac{1}{4\beta^2} + \frac{3}{8\beta^4}]}{[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}]} [\frac{\frac{\sqrt{d}a}{2}}{\sqrt{\frac{da^2}{4} - \bar{V} + 2\beta}}]^{(\frac{d-1}{2})} \tag{C.13}$$

This is what we need to prove.

For the exponential case, the arguments are similar. Let us introduce a new exponential density along the most important direction (x_1 coordinates). For other coordinates, let us use Gaussian distributions as argued in chapter 3. Since the twisting is done with $\pi_e(x) \equiv \beta e^{\beta(x+\beta)}$ in x_1 coordinate and Gaussian densities for other coordinates, the second moment $M_2(e)$ under the exponential twist is expressed as:

$$\begin{aligned} M_2(e) &= \int_{x_i \in R, i>1, \{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2/2 \geq 0\}} \cdots \int_{x_1 < -\bar{x}} \\ &\quad (\frac{f(x_1)}{\pi_e(x_1)})^2 \pi_e(x_1) dx_1 \prod_{i>1} f(x_i) dx_i \\ &= \frac{1}{(2\pi)^{\frac{d-1}{2}}} \int_{x_i \in \Theta_i} \cdots \int_{x_1 < -\bar{x}} (\frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}}}{\beta e^{\beta(x_1+\beta)}})^2 \beta e^{\beta(x_1+\beta)} e^{-\frac{1}{2}\sum_{i>1} x_i^2} \prod_{i\geq 1} dx_i \\ &= \frac{1}{\beta(2\pi)^{\frac{d+1}{2}}} \int_{x_i \in \Theta_i} \cdots \int_{x_1 < -\bar{x}} e^{-x_1^2 - \beta(x_1+\beta) - \sum_{i>1} x_i^2/2} \prod_{i\geq 1} dx_i \\ &= \frac{e^{-\beta^2}}{\beta(2\pi)^{\frac{d+1}{2}}} \int_{x_i \in \Theta_i} \cdots \int_{x_1 < -\bar{x}} e^{-x_1^2 - \beta x_1 - \sum_{i>1} x_i^2/2} \prod_{i\geq 1} dx_i \\ &= \frac{e^{-\beta^2 + \frac{\beta^2}{4}}}{\beta(2\pi)^{\frac{d+1}{2}}} \int_{x_i \in \Theta_i} \cdots \int_{x_1 < -\bar{x}} e^{-(x_1^2 + \frac{\beta}{2})^2 - \sum_{i>1} x_i^2/2} \prod_{i\geq 1} dx_i \\ &= \frac{e^{-\beta^2 + \frac{\beta^2}{4}}}{\beta(2\pi)^{\frac{d+1}{2}}} \int_{x_i \in \Theta_i} \cdots \int_{y < -\bar{x} + \frac{\beta}{2}} e^{-y^2 - \sum_{i>1} x_i^2/2} dy \prod_{i>1} dx_i \end{aligned}$$

$$= \frac{e^{-\beta^2+\frac{\beta^2}{4}}}{\sqrt{2}\beta(2\pi)^{\frac{d+1}{2}}} \int_{x_i \in \Theta_i} \cdots \int_{\bar{y} < -\sqrt{2}\bar{x} + \frac{\beta}{\sqrt{2}}} e^{-\bar{y}^2/2 - \sum_{i>1} x_i^2/2} d\bar{y} \prod_{i>1} dx_i$$

where Θ_i is $\{x_i : x_i \in Re, i > 1, \{\frac{da^2}{4} - \bar{V} - \sum_{i>1} x_i^2 \geq 0\}$.
Applying, the familiar Laplace method to the above expression gives:

$$\approx \frac{e^{-\beta^2+\frac{\beta^2}{4}}}{\sqrt{2}\beta(2\pi)^{\frac{d+1}{2}}} \int \cdots \int_{x_i \in \Theta_i} \frac{1}{\bar{Y}}[1 - \frac{1}{\bar{Y}^2} + \frac{3}{\bar{Y}^4}] e^{-\bar{Y}^2/2 - \sum_{i>1} x_i^2/2} \prod_{i>1} dx_i$$

where : $\bar{Y} = \sqrt{2}\bar{x} - \frac{\beta}{\sqrt{2}}$. Using the fact that $\bar{x}$ and β are so close together, $\bar{x}$ in the denominator can be replaced by β. This allows the fraction to go outside the above integral to give:

$$\approx \frac{e^{-\beta^2+\frac{\beta^2}{4}}}{\sqrt{2}\beta(2\pi)^{\frac{d+1}{2}}(\frac{\beta}{\sqrt{2}})}[1 - \frac{1}{(\frac{\beta}{\sqrt{2}})^2} + \frac{3}{(\frac{\beta}{\sqrt{2}})^4}] \int \cdots \int_{x_i \in \Theta_i} e^{-\bar{Y}^2/2 - \sum_{i>1} x_i^2/2} \prod_{i>1} dx_i$$

Using the formula for $\bar{Y}$ above, and the approximation formula for $\bar{x}$ and $\bar{x}^2$ in terms of β, the second moment can be further approximated by:

$$\approx \frac{e^{-\beta^2}}{\sqrt{2}\beta(2\pi)^{\frac{d+1}{2}}(\frac{\beta}{\sqrt{2}})}[1 - \frac{1}{(\frac{\beta}{\sqrt{2}})^2} + \frac{3}{(\frac{\beta}{\sqrt{2}})^4}] \int \cdots \int_{x_i \in \Theta_i} e^{-\frac{\alpha}{2}\sum_{i>1} x_i^2} \prod_{i>1} dx_i$$

where α is given by: $\alpha = 1 + \frac{\beta}{\sqrt{\frac{da^2}{4} - \bar{V}}}$. This can again be computed in the spherical coordinates to yield:

$$\begin{aligned}
M_2(e) &= \frac{e^{-\beta^2}}{\sqrt{2}\beta(2\pi)(\frac{\beta}{\sqrt{2}})}[1 - \frac{1}{(\frac{\beta}{\sqrt{2}})^2} + \frac{3}{(\frac{\beta}{\sqrt{2}})^4}](\frac{1}{\alpha})^{(\frac{d-1}{2})}\{1+\varepsilon\} \\
&= \frac{e^{-\beta^2}}{\sqrt{2}\beta(2\pi)(\frac{\beta}{\sqrt{2}})}[1 - \frac{1}{(\frac{\beta}{\sqrt{2}})^2} + \frac{3}{(\frac{\beta}{\sqrt{2}})^4}](\frac{1}{1+\frac{\beta}{\sqrt{\frac{da^2}{4} - \bar{V}}}})^{(\frac{d-1}{2})}\{1+\varepsilon\} \\
&\approx \frac{e^{-\beta^2}}{\sqrt{2}\beta(2\pi)(\frac{\beta}{\sqrt{2}})}[1 - \frac{1}{(\frac{\beta}{\sqrt{2}})^2} + \frac{3}{(\frac{\beta}{\sqrt{2}})^4}](\frac{1}{1+\frac{\beta}{\sqrt{\frac{da^2}{4} - \bar{V}}}})^{(\frac{d-1}{2})}
\end{aligned}$$

where ε is given by:

$$\varepsilon = \begin{cases} \frac{2}{\sqrt{\pi}} \int_{-\infty}^{-\sqrt{\alpha}\rho} e^{-\frac{y^2}{2}} dy + \sqrt{\frac{2}{\pi}} \alpha^{(\frac{d}{2})} \{\sum_{i=1}^{[\frac{d-1}{2}]} \frac{(\frac{d-i}{2})! 2^{(\frac{d-i}{2})}}{\alpha^i (d-i)!} \rho^{(d-2i)}\} e^{-\frac{\alpha\rho^2}{2}} & \text{if } d = odd \\ \alpha^{(\frac{d}{2})} \{1 + \sum_{i=1}^{[\frac{d-1}{2}]} \frac{1}{\alpha^i (\frac{d}{2}-i)! 2^{(\frac{d}{2}-i)}} \rho^{(d-2i)}\} e^{-\frac{\alpha\rho^2}{2}} & \text{if } d = even. \end{cases}$$

Therefore, the efficiency for the exponential twist is given by:

$$Efficiency = \frac{\sigma^2}{\sigma_M^2} = \frac{p - p^2}{M_2(e) - p^2} = \frac{1 - p}{\frac{M_2(e)}{p} - p}$$

where

$$\frac{M_2(e)}{p} = \frac{e^{-\frac{\beta^2}{2}}}{\beta\sqrt{2\pi}} \frac{[1 - \frac{4}{\beta^2} + \frac{12}{\beta^4}]}{[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}]}. \tag{C.14}$$

This is what we need to prove.

As for the proof of the estimate of efficiency as β goes infinity, one can obtain the estimate by simply putting the previous results together as follows:

$$Efficiency = \frac{\sigma^2}{\sigma_M^2} = \frac{1-p}{\frac{M_2}{p} - p} = \frac{\frac{1}{p} - 1}{\frac{M_2}{p^2} - 1} \sim \frac{\frac{1}{p}}{\frac{M_2}{p^2}} = \frac{1}{\frac{M_2}{p}}$$

Therefore, taking the limit of $\frac{M_2}{p}$ for both Gaussian and exponential twist gives:

$$\begin{aligned}
\frac{1}{\frac{M_2(g)}{p}} &= 2e^{\frac{\beta^2}{2}} \frac{[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}]}{[1 - \frac{1}{4\beta^2} + \frac{3}{8\beta^4}]} [\frac{\sqrt{\frac{da^2}{4} - \bar{V}} + 2\beta}{\frac{\sqrt{d}a}{2}}]^{(\frac{d-1}{2})} \sim Const \times e^{\frac{\beta^2}{2}} \\
\frac{1}{\frac{M_2(e)}{p}} &= \beta\sqrt{2\pi} e^{\frac{\beta^2}{2}} \frac{[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}]}{[1 - \frac{4}{\beta^2} + \frac{12}{\beta^4}]} \sim Const \times \beta e^{\frac{\beta^2}{2}}
\end{aligned} \tag{C.15}$$

The last step is justified since β, $\frac{\sqrt{d}a}{2}$, and $\sqrt{\frac{da^2}{4} - \bar{V}}$ are almost of the same order of magnitude.

QED.

The β above basically tells how far away the new mean is shifted from the original distribution. Therefore, the bigger the β, the more efficient the importance sampling would be. Furthermore, this theorem shows that the importance sampling using the exponential twisted density is more efficient than the Gaussian twist at least by a factor of β.

C3 (B_v: d-Dimensional Convex Set with Low Curvature)

So far we assumed that the VAR region is a sphere in R^d. However, we can make more powerful statement about the efficiency of the exponential twisted density for general cases using the theorem above. Without the loss of generality, let us assume that the density is uncorrelated, and B_v is convex. From the Large Deviation Theory (LDT: Chapter 3), if β is a local minimum, then the importance twist is created at β. Typically $p = P(B_v)$ is very small by the definition of VAR. Therefore, B_v is located farther away from the origin. Therefore, one can find a open set (a sphere) K such that $P(K) = P(B_v) + o(1)$. K can be chosen so that the curvature of K is the same as the curvature of B_v at β, and K touches B_v at β. Typically, B_v is shaped like a half space with some curvature at β. If the curvature of the VAR region B_v at β is very small, the center of K has to be located very far away from the origin to cover B_v and keep the same curvature as K at the same time. This implies that the radius of K must be very large. Using the same notation from the previous theorem, the radius of K can be expressed as $\sqrt{\frac{da^2}{4} - \bar{V}}$. Therefore, we have $\sqrt{\frac{da^2}{4} - \bar{V}} \gg \beta$. Thus, there exists $s > 0$ such that $\sqrt{\frac{da^2}{4} - \bar{V}} = \beta^{1+s}$. We have shown that the exponential twist is more efficient than the Gaussian twist by the order of β. We shall show that for large enough s (smaller curvature of B_v), the exponential twist exhibits the efficiency of β^5 over the Gaussian twist for this case.

Thorem 9.0.4 (General Theorem: The Efficiency for the Exponential Twisted Density)
Suppose we have the assumptions as given in the previous paragraph. If $s > 4$, the variance for the exponential twist decreases (or more efficient) faster than the Gaussian twist by a factor of β^5 for large β. To be more precise, we have:

$$Efficiency = \frac{M_2(g) - p^2}{M_2(e) - p^2} = \frac{\sqrt{2\pi}\beta^5}{10}(1 + o(\frac{1}{\beta^{min(1,s-4)}}))$$

Proof: From the previous theorem, we have the following results:

$$\begin{aligned}
p &= \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta}[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}]\frac{1}{\{1 + \frac{\beta}{\sqrt{\frac{da^2}{4} - \bar{V}}}\}^{\frac{d-1}{2}}} + o(1) \\
M_2(g) &= \frac{e^{-\beta^2}}{\sqrt{2\pi}(2\beta)}[1 - \frac{1}{(2\beta)^2} + \frac{3}{(2\beta)^4}]\frac{1}{\{1 + \frac{2\beta}{\sqrt{\frac{da^2}{4} - \bar{V}}}\}^{\frac{d-1}{2}}} + o(1) \\
M_2(e) &= \frac{e^{-\beta^2}}{2\pi\beta^2}[1 - \frac{2}{\beta^2} + \frac{12}{\beta^4}]\frac{1}{\{1 + \frac{\beta}{\sqrt{\frac{da^2}{4} - \bar{V}}}\}^{\frac{d-1}{2}}} + o(1)
\end{aligned}$$

These can be simplied as:

$$\begin{aligned}
p &= \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta}[1 - \frac{1}{\beta^2} + \frac{3}{\beta^4}]\frac{1}{\{1 + \frac{1}{\beta^s}\}^{\frac{d-1}{2}}} + o(1) \\
M_2(g) &= \frac{e^{-\beta^2}}{\sqrt{2\pi}(2\beta)}[1 - \frac{1}{(2\beta)^2} + \frac{3}{(2\beta)^4}]\frac{1}{\{1 + \frac{2}{\beta^s}\}^{\frac{d-1}{2}}} + o(1)
\end{aligned}$$

$$M_2(e) = \frac{e^{-\beta^2}}{2\pi\beta^2}[1 - \frac{2}{\beta^2} + \frac{12}{\beta^4}]\frac{1}{\{1+\frac{1}{\beta^s}\}^{\frac{d-1}{2}}} + o(1)$$

For large β, we have: $\frac{1}{\{1+\frac{1}{\beta^s}\}^{\frac{d-1}{2}}} = 1 - \frac{2}{\beta^s} + o(\frac{1}{\beta^{2s}})$. If $s > 4$, then we can reduce the following algebra as:

$$\begin{aligned} & (1 - \frac{2}{\beta^2} + \frac{12}{\beta^4}) - (1 - \frac{1}{\beta^2} + \frac{3}{\beta^4})^2(1 - \frac{2}{\beta^s} + o(\frac{1}{\beta^{2s}})) \\ = \; & 1 - \frac{2}{\beta^2} + \frac{12}{\beta^4} - (1 + \frac{1}{\beta^4} + \frac{9}{\beta^8} - \frac{2}{\beta^2} + \frac{6}{\beta^4} - \frac{6}{\beta^6} + o(\frac{1}{\beta^s})) \\ = \; & \frac{5}{\beta^4} + o(\frac{1}{\beta^s}) \end{aligned}$$

Therefore, $M_2(e) - p^2$ becomes:

$$\begin{aligned} M_2(e) - p^2 &= \frac{e^{-\beta^2}}{2\pi\beta^2}\frac{1}{\{1+\frac{1}{\beta^s}\}^{\frac{d-1}{2}}} \\ & [(1 - \frac{2}{\beta^2} + \frac{12}{\beta^4}) - (1 - \frac{1}{\beta^2} + \frac{3}{\beta^4})^2(1 - \frac{2}{\beta^s} + o(\frac{1}{\beta^{2s}}))] \\ &= \frac{e^{-\beta^2}}{2\pi\beta^2}\frac{1}{\{1+\frac{1}{\beta^s}\}^{\frac{d-1}{2}}}(\frac{5}{\beta^4} + o(\frac{1}{\beta^s})) = \frac{5e^{-\beta^2}}{2\pi\beta^6}\frac{1}{\{1+\frac{1}{\beta^s}\}^{\frac{d-1}{2}}}(1 + o(\frac{1}{\beta^{s-4}})) \end{aligned}$$

Similarly, we can compute $M_2(g) - p^2$ as:

$$\begin{aligned} M_2(g) - p^2 &= \frac{e^{-\beta^2}}{\sqrt{2\pi}\beta}\frac{1}{\{1+\frac{1}{\beta^s}\}^{\frac{d-1}{2}}} \\ & [\frac{1}{2}(1 - \frac{1}{(2\beta)^2} + \frac{3}{(2\beta)^4}) - \frac{1}{\sqrt{2\pi}\beta}(1 - \frac{1}{\beta^2} + \frac{3}{\beta^4})^2\frac{\{1+\frac{2}{\beta^s}\}^{\frac{d-1}{2}}}{\{1+\frac{1}{\beta^s}\}^{\frac{d-1}{2}}}] + o(1) \\ &= \frac{e^{-\beta^2}}{\sqrt{2\pi}\beta}\frac{1}{\{1+\frac{1}{\beta^s}\}^{\frac{d-1}{2}}}(\frac{1}{2} + o(\frac{1}{\beta})) = \frac{e^{-\beta^2}}{2\sqrt{2\pi}\beta}\frac{1}{\{1+\frac{1}{\beta^s}\}^{\frac{d-1}{2}}}(1 + o(\frac{1}{\beta})) \end{aligned}$$

Therefore, for large β we have:

$$\begin{aligned} Efficiency &= \frac{M_2(g) - p^2}{M_2(e) - p^2} = \frac{e^{-\beta^2}}{2\sqrt{2\pi}\beta}(1 + o(\frac{1}{\beta}))\frac{2\pi\beta^6}{5e^{-\beta^2}}(1 + o(\frac{1}{\beta^{s-4}})) \\ &= \frac{\sqrt{2\pi}\beta^5}{10}(1 + o(\frac{1}{\beta^{min(1,s-4)}})) \end{aligned}$$

QED.

C4 Computational Results

For illustrative purposes, several experiments have been performed to check the accuracy of the formula given in the previous section. 5 different cases are investigated. In each case, 10,000 iterations are run for portfolios of 10 independent underliers (10 dimensions). The deltas, the gammas, and the VARs are set forth as follows:

[**Table C1: Data**]

Case	Delta	Gamma	VAR
1	4.619649e+001	3.765770e+000	-3.292969e+002
2	4.649649e+001	3.765770e+000	-3.900297e+002
3	4.009649e+001	3.765770e+000	-2.539715e+002
4	4.919649e+001	3.765770e+000	-1.939715e+002
5	4.619649e+001	3.765770e+000	-2.053972e+002

The following table shows the results for the Gaussian cases. One can notice the closeness of the analytic approximations to the simulation versions.

[**Table C2: Computational Results for the Gaussian Case**]

Case	Simulated Prob.	Analytic Prob.	Simulated σ^2	Analytic σ^2
1	4.007706e-3	4.219097e-3	5.946114e-5	6.26320e-5
2	9.372176e-4	9.965419e-4	3.890290e-6	3.94848e-6
3	2.403404e-4	2.508348e-4	3.111087e-7	3.21520e-7
4	2.496546e-4	2.427197e-4	2.590840e-7	2.59308e-7
5	1.381156e-5	1.334211e-5	1.079736e-9	9.85356e-10

The next table shows the results for the exponential cases. One can similarly notice the closeness of the analytic approximations to the simulation versions. One important observation is that the variances for the exponential cases are smaller by additional factors of 4 to 17 than the variances for the Gaussian cases as expected from the theorem.

[**Table C3: Computational Results for the Exponential Case**]

Case	Simulated Prob.	Analytic Prob.	Simulated σ^2	Analytic σ^2
1	4.022585e-3	4.219097e-3	1.396714e-5	1.306192e-5
2	9.297350e-4	9.965419e-4	9.869584e-7	9.947712e-7
3	2.483174e-4	2.508348e-4	6.012005e-8	6.042611e-8
4	2.429384e-4	2.427197e-4	1.580104e-8	1.468649e-8
5	1.368510e-5	1.334211e-5	6.705168e-11	6.681958e-11

The next table illustrates how variances for each Gaussian and exponential twists become smaller as β becomes larger. It is clear that the variances for the exponential twist become smaller at much faster pace than the Gaussian twist as β becomes larger. This is to be expected from the theorem as well (in asymptotic sense).

[**Table C4: Variance Ratio Results**]

Case	Gaussian σ^2	Exponential σ^2	Variance Ratio	β
1	6.26320E-5	1.30619E-5	4.79501E+0	2.40300E+0
2	3.94848E-6	9.94771E-7	3.96923E+0	2.86250E+0
3	3.21520E-7	6.04261E-8	5.32087E+0	3.28620E+0
4	2.59308E-7	1.46865E-8	1.76562E+1	3.42170E+0
5	9.85356E-10	6.68196E-11	1.47465E+1	4.12440E+0

[APPENDIX D: Data Used for the Numerical Results in Chapter 5]

[Data Used for Table 1: '2 Dim. Case']
# of Assets = 2, Risk Free Rate = 0.07, Initial Values =[18,24]
Volatility sets for each = [.2,.18], Drift =[.09,.12], Cross correlation = 0.25.

Portfolio Characteristics:
(1) Long 150 on the first security & Long 20 Calls on the first security struck at 23, expiry=0.9 yrs
(2) Long 100 on the second security & Short 20 Puts on the second security struck at 22, expiry= 0.7 yrs

[Data Used for Table 1: '8 Dim. Case']
# of Assets = 8, Risk Free Rate = 0.07
Initial Values =[35,45,10,32,70,30,48,21]
Volatility sets for each assets =[.2,.23,.3,.2,.14,.11,.16,.21]
Drift =[.15,.09,.12,.08,.04,.1,.085,.09]
Correlation Matrix Used =

[1.0000,0.0497,0.1579,0.0648,0.0744,0.0498,0.0507,0.0583;
0.0497,1.0000,-0.0843,-0.1134,-0.1916,-0.4140,0.4857,-0.2857;
0.1579,-0.0843,1.0000, 0.1474, 0.4641,-0.0192,-0.0889,0.5782;
0.0648,-0.1134,0.1474,1.000,-0.2782,0.3582,-0.3612,0.0268;
0.0744,-0.1916,0.4641,-0.2782,1.0000,-0.192,-0.0101,0.5875;
0.0498,-0.4140,-0.0192,0.3582,-0.1920,1.0000,-0.0715,-0.0865;
0.0507,0.4857,-0.0889,-0.3612,-0.0101,-0.0715,1.0000,-0.2561;
0.0583,-0.2857,0.5782, 0.0268,0.5875,-0.0865,-0.2561,1.0000]

Portfolio Characteristics:
(1) Short 30 on the first security
(2) Long 13 on the second security & Short 400 Call on the second security struck at 49, expiry= 0.8 yrs
(3) Long 10 on the third security & Long 200 Puts on the third security struck at 12, expiry=0.9 yrs
(4) Long 21 on the fourth security & Long 10 Calls on the fourth security struck at 30, expiry=0.52 yrs
(5) Long 100 on the fifth security & Short 100 Puts on the fifth security struck at 68, expiry=1 yr
(6) Long 300 on the sixth security
(7) Long 230 on the seventh security & Short 300 Puts on the seventh security struck at 50, expiry=0.6 yrs
(8) Long 49 on the eighth security

[APPENDIX E: Computational Results]

[Tables 1. 2 Dimensional and 8 Dimensional Portfolio Examples]

p = The probability, $GHS\ Ratio$ = Variance ratio by the GHS method, $Gauss.\ Ratio$ = Variance ratio by Gaussian twist with optimization, $Exp.\ Ratio$ = Variance ratio by exponential twist with optimization, v = VAR, t= Horizon in years, Number of Iterations = 50,000. Refer to Appendix D for the detail descriptions of '2 Dim. Case' and '8 Dim. Case' portfolios.

(Test Case Scenario Description)

(1) 2 Dim. Case, v=4100, t=.3288, Extreme Quantile Long Horizon
(2) 2 Dim. Case, v=4560, t=.3288, Non-Extreme Quantile Long Horizon
(3) 2 Dim. Case, v=4825, t=.0274, Extreme Quantile Long Horizon
(4) 2 Dim. Case, v=4907, t=.0274, Non-Extreme Quantile Long Horizon
(5) 2 Dim. Case, v=5010, t=.00274, Extreme Quantile Short Horizon
(6) 2 Dim. Case, v=5035, t=.00274, Non-Extreme Quantile Short Horizon
(7) 8 Dim. Case, v=26400, t=.0822, Extreme Quantile Long Horizon
(8) 8 Dim. Case, v=27451, t=.0822, Non-Extreme Quantile Long Horizon

[Table 1.1: Importance Sampling Only]

Case	p	$GHS\ Ratio$	$Gauss.\ Ratio$	$Exp.\ Ratio$
(1)	1.0208E-02	2.28312E+01	2.49433E+01	1.26454E+02
(2)	5.0104E-02	7.12314E+00	8.07231E+00	4.84211E+01
(3)	1.0314E-02	2.35212E+01	2.98211E+01	1.39213E+02
(4)	5.1355E-02	5.91344E+00	6.12322E+00	4.02126E+01
(5)	1.1443E-02	1.92112E+01	1.93212E+01	1.20121E+02
(6)	4.9436E-02	2.10012E+00	2.03112E+00	1.32122E+01
(7)	3.7103E-03	4.98991E+01	5.12133E+01	2.23111E+02
(8)	6.7008E-02	9.08113E+00	8.99121E+00	2.01221E+01

[Table 1.2: Importance Sampling + Control Variate]

Case	p	$GHS\ Ratio$	$Gauss.\ Ratio$	$Exp.\ Ratio$
(1)	1.0203E-02	2.76823E+00	2.98322E+01	1.28212E+02
(2)	5.0111E-02	1.62123E+01	1.73212E+01	6.34222E+01
(3)	1.0268E-02	2.33312E+01	2.99222E+01	1.38211E+02
(4)	5.0999E-02	8.26311E+00	9.14878E+00	6.51211E+01
(5)	1.1312E-02	4.13221E+01	4.19211E+01	1.75212E+02
(6)	4.9521E-02	1.81211E+01	1.78211E+01	4.42113E+01
(7)	3.7914E-03	4.76665E+01	4,99121E+01	2.02121E+02
(8)	6.7123E-02	4.89112E+01	5.01213E+01	1.09312E+02

[Tables 2. 10 Dimensional Case, Quantile $\approx$ 0.01, Horizon=10 days]

(Assumptions)

Number of Iterations = 10,000
Number of Assets = 10, All Uncorrelated
Initial Values for all Assets = 100 , Volatility for all Assets = 0.3
Expiry for Portfolio (1-3) = 0.5 yrs , Expiry for Portfolio (4-8) = 0.1 yrs
Risk Free Rate = 0.05, Dividend or Growth Rate =0.0
Reverse Probability = 0.01 (99 %)
P&L Horizon = 10 days/250 trading days (t=0.04)
Barrier for DAO option = 95

"DAO" stands for Down-And-Out option, "DAI" stands for Down-And-In option, and "CON" stands for Cash-Or-Nothing option (or sometimes called "digital"). "Variance Ratio" is the ratio of variances =
Variance(Reg. MC)/Variance(Importance Sampling). Here, 'GHS Method' refers to the importance sampling method proposed by GHS(Glasserman, etc) based on analytic approximation (See Appendix B). "Optimization Method" refers to the importance sampling method using non-linear optimization method (sometimes involves multiple local minima). The word "Puts" are assumed to be plain vanilla European puts, and "Calls" are assumed to be plain vanilla calls. All options in this experiments are assumed to be European ATM (At-the-Money) unless otherwise stated. The first set of portfolio is described as follows:

(Portfolio Description)

(1) 10 Short Calls, 5 Short Puts.

(2) 10 Long Calls, 5 Long Puts.

(3) 10 Long Calls, 5 Short Puts.

(4) 10 Short Calls, Puts Delta Hedged *.

(5) 10 DAO Short Calls, 5 Short Puts.

(6) 10 DAO Short Calls, Puts Delta Hedged *.

(7) 10 DAO Short Calls, CON Puts Delta Hedged *.

(8) 10 DAO Short Calls, DAO Puts Delta Hedged *.

(9) 10 DAO Short Calls, 10 Long DAI Calls .

(* involves more than one local minimum)

[Table 2.1: Variance Ratio Test for the Gaussian Twist by the GHS Method]

Portfolio	Probability	Variance Ratio	Accepted Points
(1)	1.01533E-02	3.20013E+01	5011
(2)	1.14321E-02	3.41234E+01	5123
(3)	1.03943E-02	3.31233E+01	4926
(4)	1.08323E-02	1.75335E+01	4321
(5)	1.12342E-02	1.04843E+01	3850
(6)	9.65493E-03	9.18233E+00	2534
(7)	9.31232E-02	3.12344E-01	320
(8)	1.02123E-02	7.34232E+00	2100
(9)	1.03212E-02	3.23423E+01	4532

[Table 2.2: Variance Ratio Test for the Gaussian Twist by Non-Linear Optimization]

Portfolio	Probability	Variance Ratio	Accepted Points
(1)	1.00123E-02	3.31234E+01	5121
(2)	1.10313E-02	2.83456E+01	5322
(3)	1.04534E-02	3.12223E+01	4926
(4)	1.10234E-02	2.12234E+01	5010
(5)	1.15623E-02	1.92334E+01	4998
(6)	1.00123E-02	1.38283E+01	5200
(7)	1.10112E-02	1.21314E+01	4850
(8)	1.09101E-02	1.34212E+01	4980
(9)	1.00013E-02	3.81234E+01	5032

[Table 2.3: Variance Ratio Test for the Exponential Twist by Non-Linear Optimization]

Portfolio	Probability	Variance Ratio	Accepted Points
(1)	1.01122E-02	1.29874E+02	9558
(2)	1.04315E-02	1.31235E+02	9856
(3)	1.03531E-02	4.35323E+02	9983
(4)	1.07843E-02	4.65234E+01	9001
(5)	1.11232E-02	4.21433E+01	8992
(6)	1.00834E-02	4.87683E+01	8999
(7)	1.09945E-02	3.21314E+01	9010
(8)	1.08934E-02	3.94342E+01	8934
(9)	1.00023E-02	1.12434E+03	9993

[Tables 3. 100 Dimensional Case, Quantile $\approx$ 0.01, Horizon=10 days]

(Assumptions)

The parameters and conditions are the same as Tables 9, except: Number of Assets = 100, All Uncorrelated, Expiry for Portfolio = 0.1 yrs

(Portfolio Description)

(1) 10 Short Calls, 10 Short Puts.

(2) 10 Long Calls, 10 Long Puts.

(3) 10 Long Calls, 5 Short Puts.

(4) 10 Short Calls, Puts Delta Hedged *.

(* involves more than one local minimum)

[Table 3.1: Variance Ratio Test for the Gaussian Twist by the GHS Method]

Portfolio	Probability	Variance Ratio	Accepted Points
(1)	1.01244E-02	1.89234E+01	5009
(2)	1.09234E-02	2.40233E+01	5113
(3)	1.05642E-02	2.23456E+01	5120
(4)	1.13553E-02	1.09345E+01	4820

[Table 3.2: Variance Ratio Test for the Gaussian Twist by Non-Linear Optimization]

Portfolio	Probability	Variance Ratio	Accepted Points
(1)	1.00123E-02	2.12134E+01	5121
(2)	1.10313E-02	2.54413E+01	5322
(3)	1.04534E-02	3.98753E+01	4926
(4)	1.10235E-02	1.15234E+01	5010

[Table 3.3: Variance Ratio Test for the Exponential Twist by Non-Linear Optimization]

Portfolio	Probability	Variance Ratio	Accepted Points
(1)	1.00122E-02	2.12131E+02	9885
(2)	1.10314E-02	2.54414E+02	9832
(3)	1.04536E-02	7.69725E+02	9983
(4)	1.10233E-02	3.15342E+01	8912

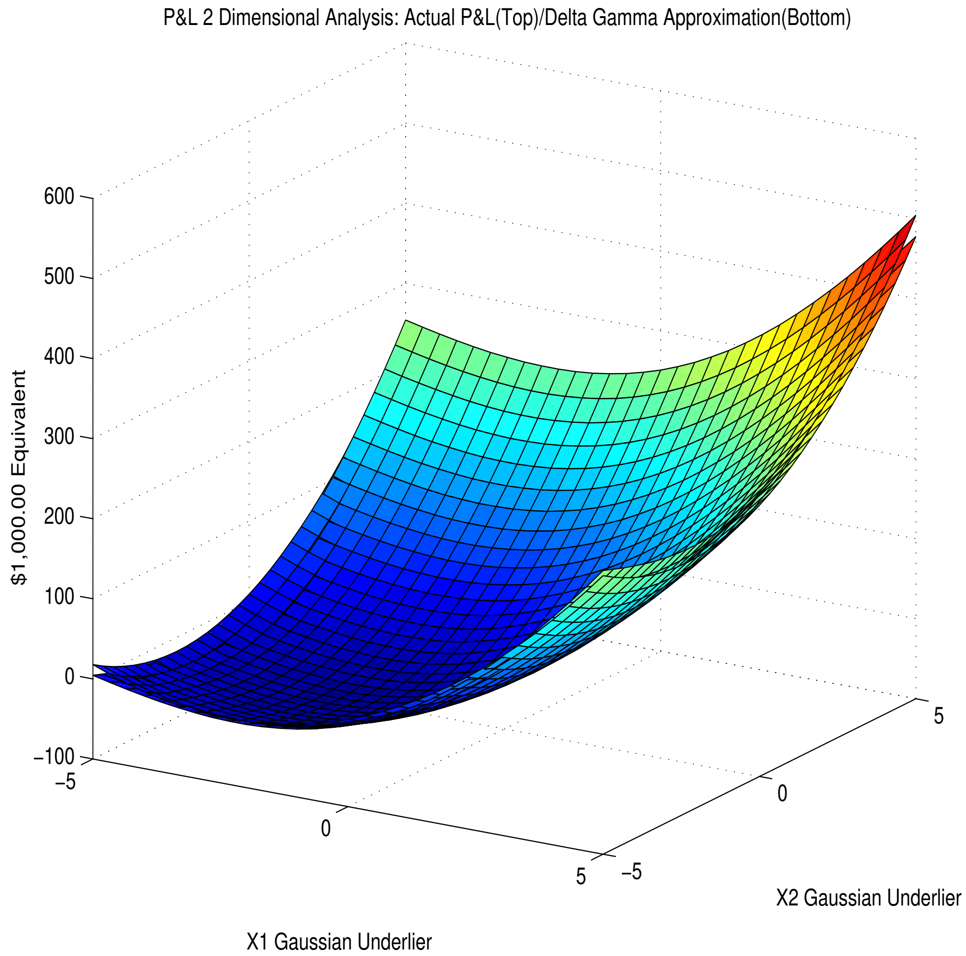

Figure 9.2: P&L (the top graph) of a portfolio containing 10 long European calls and 5 long European puts in 2 dimensions, and the analytic approximation (the bottom graph) for the same portfolio. Note that in this case, the approximation is so excellent that these two graphs appears to be almost identical. Numerical tests also confirm that the probability for the actual P&L and that for the analytic approximation are very close. One can see from the graph that the extreme event can be effectively sampled from a new density which is centered around (-2,-2).

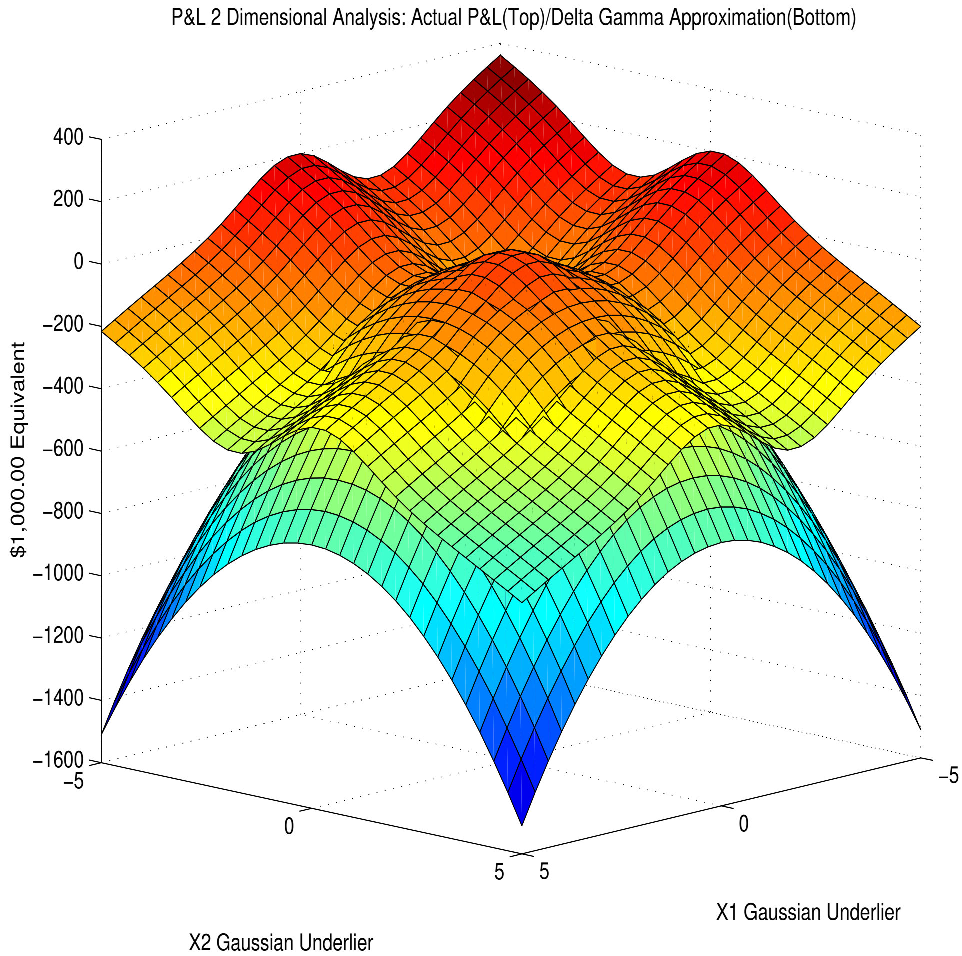

Figure 9.3: P&L (the top graph) of a portfolio containing 10 short European DAO (Down-And-Out) calls delta hedged with DAO puts in 2 dimensions, and the analytic approximation (the bottom graph) for the same portfolio. The analytic approximation is so far off that sampling based on this approximation becomes no longer helpful. However, one can see visually that the extreme event can be sampled efficiently somewhere around (2.6,2.6). Numerical experiments demonstrates that there is an improvement in the variance using new density centered around (2.6,2.6).

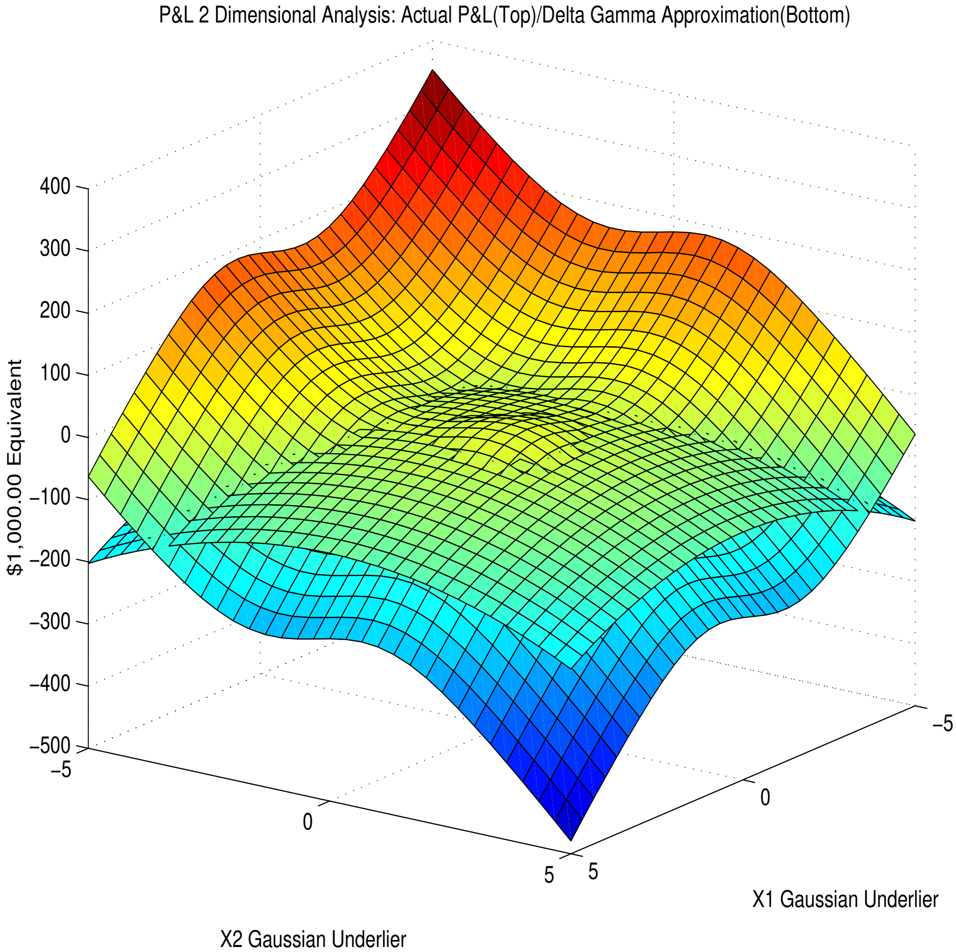

Figure 9.4: P&L (the top graph) of a portfolio containing 10 short European DAO (Down-And-Out) calls delta hedged with European Cash-Or-Nothing (Digital) puts in 2 dimensions, and the analytic approximation (the bottom graph) for the same portfolio. The Analytic approximation is so far off from the actual P&L distribution also. The graph indicate that sampling from three points may be necessary at somewhere near(-2.5,-2.5), near (-4,0), and (0,-4).

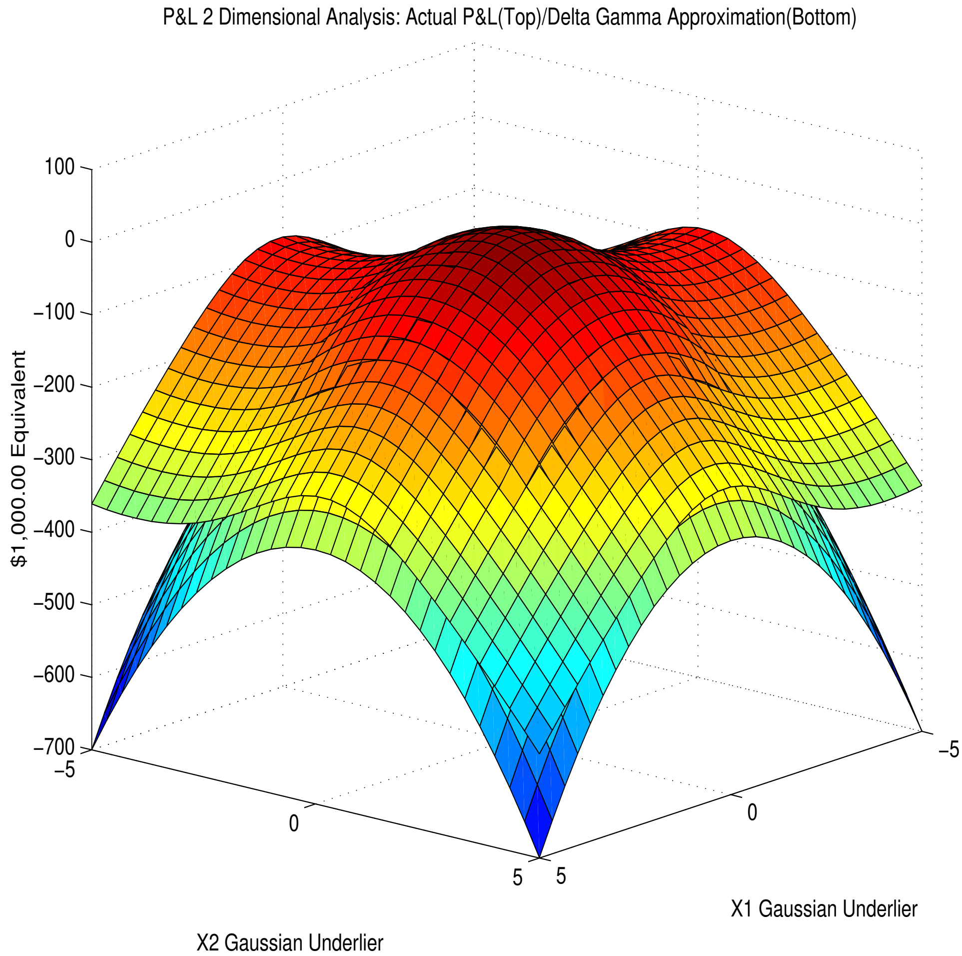

Figure 9.5: P&L (the top graph) of a portfolio containing 10 short European DAO (Down-And-Out) calls delta hedged with European puts in 2 dimensions, and the analytic approximation (the bottom graph) for the same portfolio. In this example, the analytic approximation is somewhat closer to the actual P&L. However, the graph indicates that one needs to sample from many densities with shifted means since there are countably many points that can qualify as new shifted means.

Bibliography

[1] C. Andrieu, E. Moulines, and P. Priouret, "Stability of Stochastic Approximation under Verifiable Condition", Preprint, 2002.

[2] C. Andrieu, and C.P. Robert, "Controlled MCMC for Optimal Sampling", Preprint, 2001.

[3] Dimitri P. Bersekas, Constrained Optimization and Lagrange Multiplier Methods, Academic Press, New York, 1982.

[4] M. Britten-Jones, and S.M. Schaefer, "Non-linear Value-at-Risk", European Finance review, 2:161-187, 1999.

[5] James Bucklew, Large Deviation Techniques in Decision, Simulation, and Estimation, John Wiley, New York, 1990.

[6] James. A. Bucklew, and John Sadowsky, "On Large Deviations Theory and Asymptotically Efficient Monte Carlo Estimation", IEEE Transactions on Information Theory, Vol. 36, No. 3, May 1990.

[7] James. A. Bucklew, P. New, and John S. Sadowsky, "Monte Carlo Simulation and Large Deviations Theory for Uniformly Reccurent Markov Chains", J. Applied Prob., Mar. 1990.

[8] Juan Cardenas, Emmanuel Fruchard, Etienne Koehler, Christophe Michel, and Isabelle Thomazeau, "VaR: ONE STEP BEYOND", RISK, VOL 10, NO 10, 1997.

[9] Juan Cardenas, Emmanuel Fruchard, Jean-Francois Picron, Cecilia Reyes, Kristen Walters, Weiming Yang, "Monte Carlo within a day", RISK, February, Vol 12. No 2, 1999.

[10] H. Chen, L. Guo, and A. Gao, "Convergence and Robustness of the Robbins-Monro Algorithm Truncated at Randomly Varying Bounds", Stochastic Processes and their Applications: 27, pp. 217–231, 1988.

[11] H. Chen and K. Uosaki, "Convergence of Dynamic Stochastic Approximation", Systems and Control Letters: 35, pp. 309–315, 1998.

[12] H. F. Chen and Y. M. Zhu, "Stochastic Approximation Procedures with Randomly Varying Truncations", Scientia Sinica 1, 29 (1986), pp. 914–926.

[13] H. Haario, E. Saksman, and J. Tamminen, "An Adaptive Mtropolis Algorithm", Bernoulli, 7, 2001, pp. 223–242.

[14] Banard Delyon, "Stochastic Approximation with Decreasing Gain: Convergence and Asymptotic Theory", Unpublished Lecture Notes, Universit'e de Rennes, 2000.

[15] Banard Delyon, "General Results on the Convergence of Stochastic Algorithm", IEEE Transactions on Industrial Electronics, Volume 41, Number 9, 1996.

[16] B. Delyon, M. Lavielle, and E. Moulines. "Convergence of a Stochastic Approximation Version of the EM Algorithm.", Annals of Statistics, 27 (1999), pp. 94–128.

[17] Amir Dembo, and Ofer Zeitouni, Large Deviation Techniques and Applications, Johes and Bartlett Publishers, Inc., London, 1992.

[18] S.L.S. Jacoby, J.S.Kowalik, and J.T.Pizzo, Iterative Methods for Monlinear Optimization Problems, Prentice Hall, Englewood Cliffs, New Jersey, 1972.

[19] S. Jarner and E. Hansen, "Geometric Ergodicity of Metropolis Algorithms", Stochastic Proceses and Their Applications, 85 (2000), pp. 341–361.

[20] P. Jorion, "Risk2: Measuring the Risk in Value-At-Risk," Financial Analysts Journal 52 (November 1996): 47–56.

[21] P. Jorion, VaR: A Risk Compilation, Risk Publications, London, 1997.

[22] Bruno Dupire, Monte Carlo, Risk, London, 1998.

[23] Paul Embrechts, Claudia Kluppelberg, and Thomas Mikosch, Modelling Extremal Events, Springer-Verlag, Berlin, 1997.

[24] E.T. Copson, Asymptotic Expansions, Cambridge University Press, 1965.

[25] D.R. Cox and Barndorff-Nielsen, O.E., Asymptotic Techniques for Use in Statistics, Chapman and Hall, 1989.

[26] G. S. Fisherman. Monte Carlo Concepts, Algorithms, and Applications. Springer, New York, 1996.

[27] G. Fort, E. Moulines, G.O. Roberts, and J.S. Rosenthal, "On the Geometric Ergodicity of Hybrid Samplers", Working Paper (J. Appl. Prob., to appear.), 2001.

[28] Dani Gamerman, Markov Chain Monte Carlo, Chapman & Hall, England, 1997.

[29] S. Geman, and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 6, pp. 721-741, 1984.

[30] W.R. Gilks, S.R. Richardson, and D.J. Spiegelhalter, Markov Chain Monte Carlo in Practice, 1996.

[31] P.E.Gill and W. Murry, Numerical Methods for Constrained Optimization, Academic Press, London, 1974.

[32] Douglas Glass, "Importance Sampling Applied to Value at Risk", Master's Thesis Paper, New York University, 1999.

[33] Paul Glasserman, P. Heidelberger, and P. Shahabuddin, "Asymptotically Optimal Importance Sampling and Stratification for Pricing Path-Dependent Options", Mathematical Finance, Vol. 9, No. 2, April 1999, 117-152.

[34] Paul Glasserman, P. Heidelberger, and P. Shahabuddin, "Importance sampling in the Heath-Jarrow-Morton Framework", Journal of Derivatives 7, No. 1, 32-50, 1999.

[35] Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, "Importance sampling and stratification for value-at-risk", Computational Finance 1999, Y.S. Abu-Mostafa, B. LeBanon.

[36] Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, "Stratification Issues in Estimating Value-at-Risk", Proceeding of the 1999 Winter Simulation Conference, 1999.

[37] Paul Glasserman, Yashan Wang, "Counterexamples in Importance Sampling for Large Deviations Probabilities", The Annals of Applied Probability, Vol. 7, No. 3, pp. 731-746, 1997.

[38] Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, "Variance Reduction Techniques for Value-at-Risk with Heavy-Tailed Risk Factors", IBM Research Report RC 21816, Yorktown Heights, NY., 2000.

[39] Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, "Portfolio Value-at-Risk with Heavy-Tailed Risk Factors", IBM Research Report RC 21817, Yorktown Heights, NY, 2000.

[40] Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, "Efficient Monte Carlo Methods for Value-at-Risk", IBM Research Report RC 21723, Yorktown Heights, NY, 2000.

[41] Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, "Variance Reduction Techniques for Estimating Value-at-Risk", Working Paper, Columbia University, 2000.

[42] Paul Glasserman, Phillip Heidelberger, and Perwez Shahabuddin, "Gaussian Importance sampling & Stratification: Computational Issuees", Working Paper.

[43] Jonathan Goodman, and A. D. Sokal, "Multigrid Monte Carlo Method, Conceptual Foundations", Physical Review D, Vol. 40, No. 6, pp. 2035–2071, 1989.

[44] P. Hall and C. C. Heyde, Martingale Limit Theory and its Application, Academic Press Inc., New York, 1980.

[45] J.M. Hammersley and D.C. Handscomb, Monte Carlo Methods, Methuen, London, 1964.

[46] W. Hastings, "Monte Carlo Sampling Methods Using Markov Chains an Their Applications", Biometrika, 57 (1970), pp. 97–109.

[47] J. Hull, Options, Futures, and Other Derivative Securities, 2nd Edition, Prentice Hall, 1993.

[48] I. Karatzas and S.E. Shreve, Brownian Motion and Stochastic Calculus,Springer-Verlag, New York, 1988.

[49] T. Kato, Perturbation Theory for Linear Operators, Springer-Verlag, 1966.

[50] M. Kendall and A. Stuart, Advanced Theory of Statistics, Charles Griffin & Company, London, 1977.

[51] P.E. Kloaden and E. Platen, Numerical Solution of Stochastic Differential Equations, 2nd ed., Springer-Verlag, New York, 1995.

[52] H. Kushner and G. Yin, Stochastic Approximation Algorithms and Applications, SpringerVerlag, NewYork, 1997.

[53] A.W. Lo, and A.S. Weigend, eds, MIT Press, Cambridge, Mass Jerome Barraquand, and Didier Martineau, "Numerical Valuation of High Dimentional Multivariate American Securities", Digital Paris Research Laboratory, No. 38, 1994.

[54] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equations of State Calculations by Fast Computing Machines", J. Chemical Physics, 21 (1953), pp. 1087–1091.

[55] S. Meyn and R. Tweedie, "Computable Bounds for Convergence Rates of Markov Chains.", Annals of Applied Probability, 4 (1994), pp. 981–1011.

[56] Jorge Nocedal, Stephen J. Wright, Numerical Optimization, Springer-Verlag, New York, 1999.

[57] E. Picoult, "Calculating Value-at-Risk with Monte Carlo Simulation", Monte Carlo: Methodologies and Applications for Pricing and Risk Management, B. Dupire, ed., Risk Publications, London, pp 209-229, 1999.

[58] Donald Pierre, Optimization Theory with Applications, John Wiley, New York, 1969.

[59] Elijah Polak, Optimization, Springer-Verlag, New York, 1997.

[60] RiskMetrics Group, RiskMetrics Technical Document, www.riskmetrics.com/research/techdocs, 1996.

[61] Roberts, G.O. and J.S. Rosenthal, "Geometric Ergodicity and Hybrid Markov Chains", Electronic Comm. Prob., 2:13-25, 1997.

[62] Roberts, G.O. and Tweedie, R.L., "Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithm", Biometrika 83, pp. 95–110, 1996.

[63] R. Rouvinez, "Going Greek with VaR", RISK, Feb Issue, 1997.

[64] Alan D. Sokal, "Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithm", Unpublished Lecture Notes at New York University, 1996.

[65] D. W. Strook, An Introduction to the Theory of Large Deviations, Springer-Verlag, Berlin, 1984.

[66] L. Tierney, "Markov Chains for Exploring Posterior Distributions", Ann. Statist., 22(4), pp 1701–1762, 1994.

[67] S.R.S. Varadhan, Large Deviations and Applications, Philadelphia, Pa. : Society for Industrial and Applied Mathematics, 1984.

[68] Peter Walters, Ergodic Theory - Introductory Lectures, Springer-Verlag, Berlin, 1975.

[69] Whitlock, P.A, M.H.Kalos, Monte Carlo Methods: Volume 1, John Wiley & Sons, New York, 1986.

[70] Wilmott,P., J.N. Dewynne, and D.D. Howison, Option Pricing: Mathematical Models and Computation, Oxford Financial Press, Oxford, 1993.

[71] Wismer, A. David, Chattergy, R., Introduction to Nonlinear Optimization, North-Holland, New York, 1978.

[72] Wong, R., Asymptotic Approximations of Integrals., SIAM, 1989.